Yelp Review Analysis for Bakery Shop in Santa Barbara - Group #9

1. Introduction

Yelp is an Internet company founded in 2004 to "help people find great local businesses" by providing a platform for users to write reviews of businesses. Apart from using Yelp App to find delicious food and great service as users, we data science students hope to get some advice for the business owners through data analysis, which will help develop the quality of the businesses.

In order to give more specific advice from various kinds of data in Yelp App, we only choose one type of business in a certain area: Bakery shops in Santa Barbara. The reason why we choose them is that bakeries are popular in the USA and its related dataset in Santa Barbara has more than 11,000 reviews for around 70 bakery shops in recent 10 years, which is large enough for our analysis.

To figure out the impact factors of bakery shops in Santa Barbara, we seperated the dataset into two groups: open and closed. To compare these two groups, we do exploratory data analysis (word cloud, bakery shop distribution map in Santa Barbara, etc.) to extract information from reviews and see the impact of shop distribution.

1.1 Data Sources

- Businesses Information and User Reviews – Yelp
- Average Incomes by Zip Code data – US Census Bureau,

1.2 Goal

The overall goal of this project is to provide useful, analytical insights to business owners on Yelp, and build a RShiny App to visualize our analysis, making it easy to be understood by business owners. Furthermore, we focus on this specific goal: We aim to provide advice for bakery shop owners following the key impact factors extracted from reviews and the correlation with neighbors' income. We are going to answer: "What factors might contribute to the opening and closing of bakery shops in Santa Barbara?"

2. Data Pre-Processing

Since our goal is to provide advice for bakery shops in Santa Barbara, we extract the data related and organize it into several csv files and folders: reviews-data (separated by closed or opened, and positive or negative reviews), census-data, and county_trips.csv.

2.1. Data Cleaning

The reviews can be classified to "positive review" and "negative review" by "stars" given by the customer who wrote reviews. The review with stars higher than 3.5 is classified as a positive review, otherwise it is a negative review.

Based on the reviews information, we splitted the text into tokens and retained the "business_id" and "review_id" columns for further analysis. The punctuations and stop words are stripped to ensure the pureness of the data set.

2.2. Aspect dictionary

Overviewing all these words splitted from reviews, we extracted 4 dimensions evaluating a bakery store: "food", "drink", "ambiance" and "service". We customized four dictionaries based on the reviews. The "food" and "drink" dictionaries primarily consist of neutral nouns typically classified as vocabulary related to food and beverages, without emotional preference. The other two dictionary(ie. "service" and "ambiance") are consist of words revealing the feeling of customers, like: ”cute”,”lovely” and "disappointed", and are almost all adjectives.

dictionary	food	drink	ambiance	service
Words	croissant	coffee	cute	friendly
	cake	latte	beautiful	warm
	benedict	vanilla	dirty	awful

Table 1. simply exhibition of four dimensions dictionaries

3. Exploratory Data Analysis

3.1. Word Analysis

In order to figure out the reason for store closure and give advice to those open stores, we try to do word analysis on reviews by creating word clouds. Word cloud is a useful tool for text analysis, as it visualizes words according to their appearing frequency. Word cloud also provides a function of comparing the difference between two text data sets, which is suitable for sentiment analysis.

The following are the word cloud based on the negative reviews of closed stores and the word cloud based on positive reviews of open stores, respectively. In order to compare stores in four aspects, we made comparison word clouds with four dimensions to make the contrast clear.

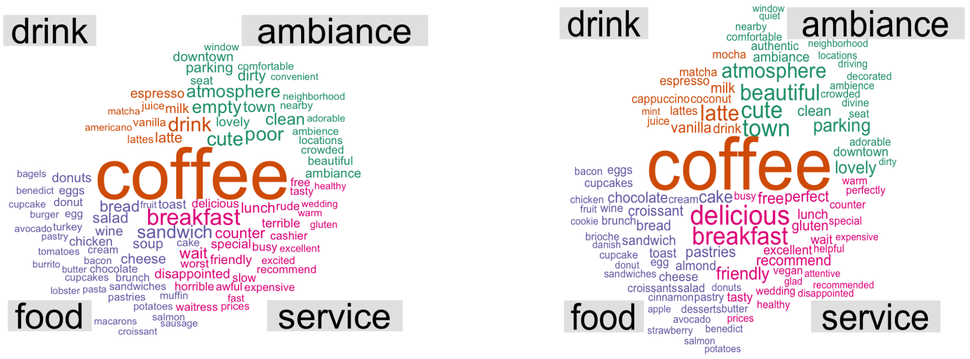

Figure 1. Negative-close (left) vs Positive-open (right)

Recognizing the significant difference of word distribution in four classifications between these two word clouds, we made another four word clouds respectively exhibiting the difference clearly and found the strength and weakness of stores in each aspect.

3.2. Sentiment analysis

Considering customers may be unwilling to show their negative feelings directly on rating stars, the review text can contain more sentiment information towards stores.

There is a sentiment lexicon inside the tidytext package of R called “afinn”, which assigns words with a score that runs between -5 to 5. Negative scores represent negative sentiment, while positive scores indicate positive sentiment. A higher score corresponds to a more positive word. We used it to measure the emotion inside the reviews and therefore calculated a sentiment score of each bakery store, which is the average sentiment score of all reviews.

3.3. Map : Income-Zipcode analysis

We draw maps in Santa Barbara, using different shades of green to mark the average income of neighbors in different zip-code areas, and point out the distribution of bakery shops with their status (open/close) from 2011 to 2021. Comparing these maps, we can speculate the impact of a neighbor's average income on bakery shop businesses. Figure 2. is an example that compares income zip-code maps in 2013 and 2019 .

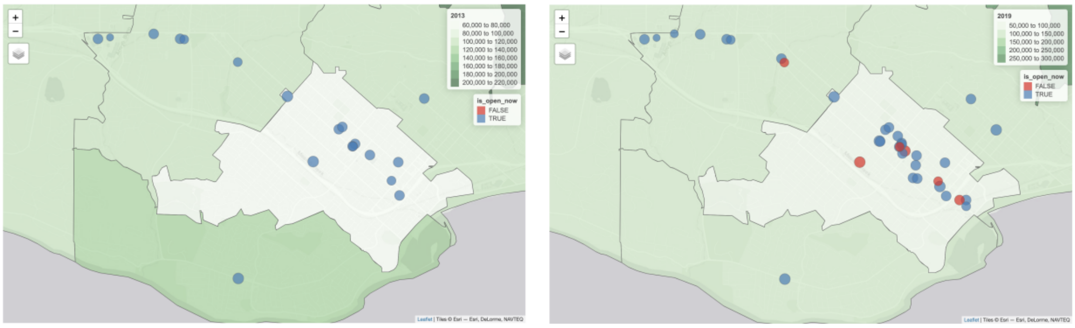

Figure 2. Income zip-code map in 2013(left) vs. Income zip-code map in 2019(right)

From these two maps, we can see the changes that happened in Santa Barbara between 2013 to 2019. First, there are more blue points on the map in 2019, which means there were more bakery shops newly opening. Second, some original blue points turn red on the map in 2019 while they were blue on the map in 2013. This means these bakery shops had been open in 2013 but closed in 2019. Clicking on a data point on each map gives information about the selected store.

4. Key findings

4.1. Word analysis

4.1.1. Food and Drink supplement

From the contrast in the word cloud, we had some important findings on the store's food and drink supplements. Firstly, coffee is one of the most important products for a bakery, as it was mentioned most frequently on the reviews. Compared with closed stores, the open stores seem to have more choice on drinks for customers, including flavors, coffee types and milk options.

For food supplements, opening stores are likely to have a good reputation on their traditional bakery items, such as cake, croissants and cupcakes. According to that, our opinion on "how to make a bakery store open for longer" includes that the owners should devote themselves to making ordinary pastries, rather than brunch items.

4.1.2. Ambiance and Service

Discussing the ambiance of stores, we also considered the location and equipment. We discovered that people often prefer visiting bakeries in their local vicinity rather than undertaking lengthy drives to visit a particular store. So we recommend someone who intends to open a bakery setting the location in a residential area. Besides, a comfortable environment and the parking space also help for the survival of bakeries.

The customers of bakeries actually concern the taste of food most, compared with other service elements. Bakery providing gluten free food will be highly praised by customers, so we consider that is a useful suggestion for potential and existing bakeries owners.

4.2. Sentiment Analysis

The average value of sentiment scores on all bakeries is 0.7, no matter if the store is closed or opening. The highest one is 2.36, which means this bakery is quite popular and praised by customers. The average score of all closed stores is 0.673, and the average score of all open stores is 0.679. Stores with scores above 1.5 are labeled as "The Best!," those with scores between 0.7 and 1.5 are labeled as "Worth a visit." Stores scoring below 0 are marked as "Many faults" while others are categorized as "Not bad", as seen in Table 2. These help the business owners get an overview on what the customers felt about the businesses.

business_id	average_sentiment_value	sentiment_label
9YFeX5sbn785n31iR3I77Q	1.774193548	The Best!
9uJuBPIQyrC0v9pVv3JKRQ	-0.023715415	Many faults
B1yPXzIIJrLWyTcE4jsYuQ	0.021341463	Not bad
8JzURExziDYuCrd7Ve9nDQ	0.823529412	Worth a visit

Table 2. Sentiment Labels based on overall review sentiment score

4.3. Aspect Score Sentiment Analysis

4.3.1. Radar plots

Reviews to all bakeries were tokenized using the Python NLTK module. Stop words are eliminated in the process. The dictionaries generated from the preliminary word-cloud analysis are used to generate individual scores for all words per aspect. Then these scores are averaged out across all bakeries to get a mean score per store. This makes the input data set for the radar plots in Figure 3. The nodes of the grey polygon represent the cumulative average scores of all the bakeries in Santa Barbara, CA for each aspect. Similarly, the nodes of the purple outlined polygon highlight the aspect scores of a selected bakery store. To read the radar plot, it is only needed to check if the nodes of the purple polygon are present inside/outside of the grey polygon to see if the selected store is performing less than average or better than average.

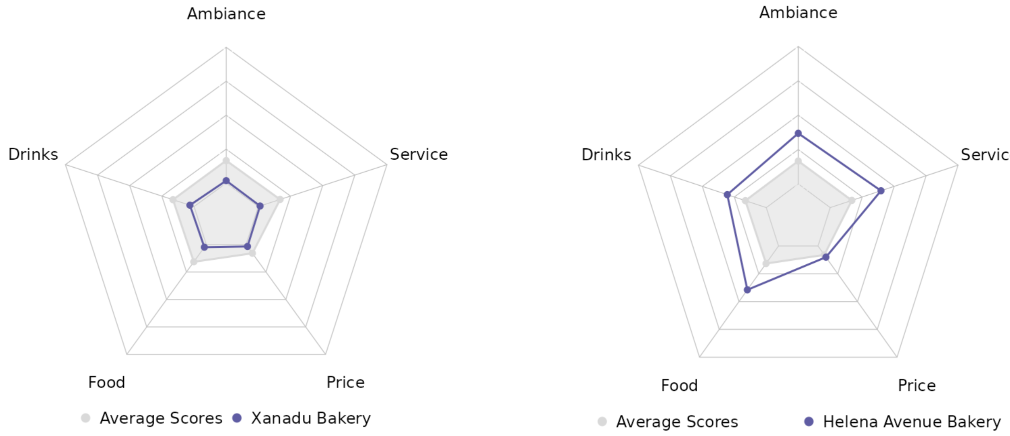

Figure 3. Comparing Aspect scores of a closed (left) and an open (right) bakery with cumulative average score

4.3.2. Star Rating prediction

The aspect scores of every bakery per star rating are calculated. Ratings are treated as factors, and the aspects are treated as independent variables. They are fitted in a Naive Bayes Classification Model. The model calculates the weight of each aspect for predicting ratings.

1 Unit Change in…	Ambience ↑	Drinks ↑	Food ↑	Price ↑	Service ↑
Open Bakeries	0.32 ↑	0.22 ↑	0.10 ↓	0.1 ↓	0.32 ↑
Closed Bakeries	0.72 ↑	0.70 ↑	0.50 ↑	1.46 ↑	0.08 ↓

Table 3. Estimated rating changes per unit change across Aspects

5. Contribution

Section/Task	Niharika Chunduru	Siyan Wang	Yiyuan Li
EDA	Income Zipcode Analysis	Word-cloud Sentiment Analysis	County Trips Analysis
Summary and Presentation	Key findings of Income Zip-code analysis, and Aspect Sentiment Analysis (Map and Radar, Review	EDA, Key findings of Sentiment Analysis (word-cloud slide and data processing slide), Review	Introduction, Data Pre-Processing, EDA, Review
Shiny App	Coding and deployment	Sentiment Labels in Map	-
Other	Pre-Processing dataset	Creating review words dictionary for aspects	County Trips Analysis