Executive Summary

Group 1: Ruofeng Tang, Bingyan Liang, Ziming Li

1 Background & Introduction

As the largest review site in the United State, Yelp provides invaluable information to business owners such as star ratings, user reviews, etc. In this project, we want to provide Mexican restaurant owners in Hillsborough, FL with recommendations that can improve their star ratings by analyzing data from Yelp, the US Department of Transportation and US Census Bureau.

2 Dataset & Preprocess

Our dataset comes from three sources. The main source is Yelp[1], which provides data rows about businesses and reviews in JSON format. Based on unique business IDs, we extracted relevant businesses (Mexican restaurants) of the relevant region (Hillsborough County, FL) from the Businesses dataset, which included 319 Mexican restaurants. The 'attributes' column included lists of restaurant attributes, so we flattened the column to multiple columns, each indicating the status of one attribute. With Google Map API, we applied reverse geocoding to get restaurants' zip codes from coordinates.

We applied the same filtering process to the Reviews dataset, and got over 30,000 reviews for Mexican Restaurants in Hillsborough County, FL. We determined the language of reviews using the 'detect' method of package 'langdetect'. The result was that there were only 101 non-English comments in total. These comments were only a small portion of our data, so they were deleted. After processing both datasets, we merged the Businesses dataset and Reviews dataset based on unique restaurant IDs.

The second source is the U.S. Bureau of Transportation Statistics and its Trips by Distance dataset[2]. We filtered out daily information about how many trips were made, how many people traveled and how many people stayed at home in Hillsborough County from 2019 to 2023. County level is the smallest module in this dataset, therefore we chose Hillsborough County even though most of Yelp Businesses data is from Tampa City. We calculated the percentage of people traveling per day for further analysis.

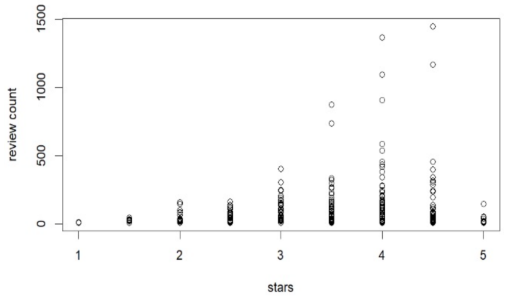
The last source is the US census Bureau[3]. Although it provided detailed information in a variety of fields, it did not provide region by region comparison tables. As a result, we manually recorded demographic information of each zip code region that included a Mexican Restaurant in Hillsborough County, FL. Our assumption is that the large Latino population in FL will influence the performance of Mexican restaurants. We merged this dataset to the Businesses dataset for further analysis.

3 Analysis Results

3.1 Analysis on Business dataset

This plot (restaurant's star rating vs. its review count) appeared right-skewed, suggesting that more restaurants had high star ratings and high review counts.

Figure 1



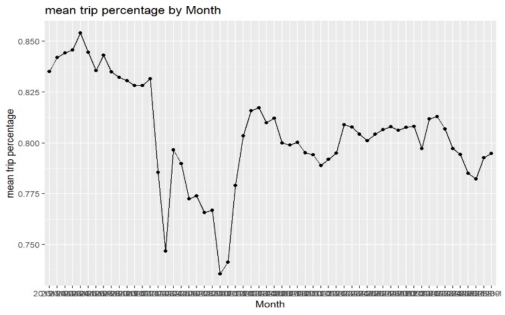
We also performed two-sample t-tests on most common attributes of Mexican restaurants, checking if they would influence star ratings. There are 11 attributes with over 100 records, so the Bonferroni correction for alpha=0.05 gives adjusted alpha=0.0045. Two attributes were significant: do not provide a TV (p value=2.7e-05, star rating difference=0.58) and do not provide a delivery option (p=4.9e-06, star rating difference=0.47). Both suggested the removal of a service would increase star rating, so one possible reason was that extra services may cause extra problems to complain about in reviews. Also, this small dataset contains only 329 rows, causing the results to be not as significant as later analyses.

With demographic data from the US Census Bureau, we concluded that the zip-code-wise Latino population did not influence the star ratings of Mexican restaurants. Our linear regression did not produce interpretable results, and no correlations were detected. However, we identified zip code regions 33603, 33604, 33614 in Tampa City, 33570 in Sun City as the best regions for Mexican restaurants. The average star rating for each region is above 4, and there are a significant number of restaurants in each of them.

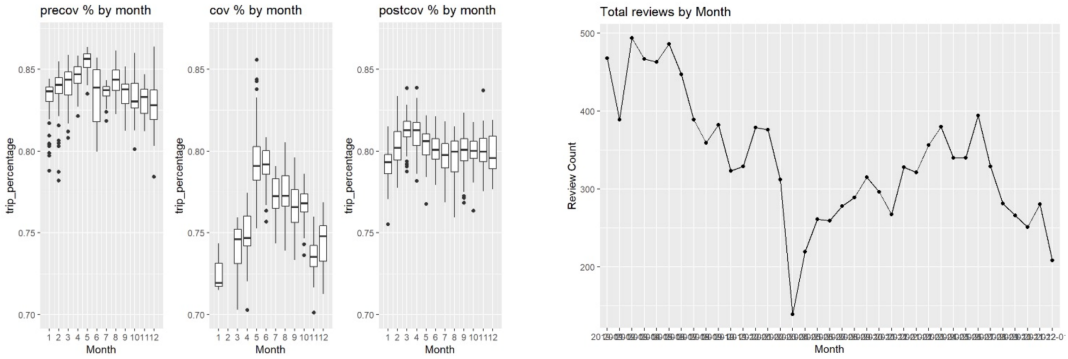
3.2 Analysis on Trips by Distance Dataset

There is a clear Covid influence on the percentage of people traveling per day. After inspection, we divided the dataset to pre-Covid, mid-Covid, post-Covid, and explored the differences. The breakpoints are Mar.13,2020 and Jan.4,2021.

Figure 2



The comparison graph below displays the percentage of people traveling per month before, during and after Covid. There was a clear drop during Covid, and it did not recover after Covid. A similar trend happened to the number of reviews on Yelp: people gave fewer reviews on Yelp during Covid, and it did not recover post Covid. We discovered other interesting trends about trips made during the whole time period, but they were less helpful to business owners. Other results were omitted.



3.3 Analysis on Reviews dataset

We produced n-grams and performed sentiment analysis on reviews using the NLTK package in Python. Then, we analyzed the reviews using the n-gram model with Latent Dirichlet Allocation (LDA) and explored the different topics associated with high and low ratings through sentiment analysis. One notable bigram was 'sour cream': it was the third most frequent bigram in reviews with bottom quantile star rating and sentiment score, while this bigram did not appear in top 20 frequent bigram in all reviews, suggesting a strong negative link.

3.3.1 Highly rated topics(1-gram)

Because they are chosen from high ratings, the topics are about positive aspects, such as "delicious food" or love for the restaurants in general; the average sentiment scores are high; taking topic 0 (quality of food) as an example, the emotion score is as high as 0.836, indicating strong positive sentiment.

3.3.2 Low rated topics(1-gram)

Figure 3

'Chipotle' , delays and order issues are always mentioned in the negative reviews; sentiment scores are low or negative, even reaching -0.219 in topic 1 (Ordering experience and customer service), reflecting dissatisfaction.

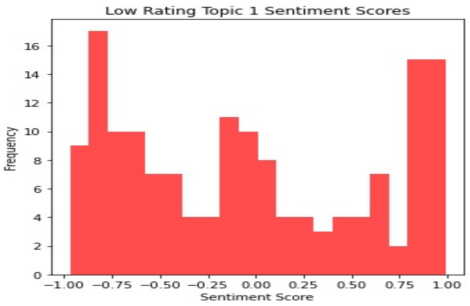
3.3.3 2-gram and 3-gram topic analysis

Similar to the results for the 1-gram, the highly rated topics focus on the quality of the food. The low-rated topics still mentioned "Chipotle" and operational issues (order, lines, services); for sentiment scores, highly rated topics generally have higher sentiment scores, while the sentiment distribution of low-rated topics is uneven(Figure 4, for example). This may indicate that there are also some positive comments among some low-rated comments, but only for specific aspects, such as food-related.

3.3.4 Statistical Significance and Model fit

Sentiment scores show clear differences between high and low ratings on Yelp's comments, but looking at the sentiment score distribution, there are some positive aspects to even low ratings. Analyzing from the perplexity value, the high perplexity value indicates that the scope of the topics is broad, which may be due to the complexity of the dataset.

Figure 4



3.3.5 Limitation

Our analysis relies on the accuracy of LDA and sentiment analysis, which may not fully capture the nuances between languages and context(eg. Sarcasm, puns); the topics of LDA need to be interpreted subjectively.

4 Recommendations to Business Owners

Based on the analysis part, we came up with some realistic recommendations: try to open the restaurant in zip code region 33603, 33604, 33614 in Tampa City or 33570 in Sun City; do not provide a TV; consider the fact that providing a delivery service would reduce average star rating on Yelp, even if a delivery service might be beneficial; be careful with the quality of sour cream; Improve the ordering process to reduce waiting times ad better train staff, especially managers. Since "Chipotle" has a higher frequency in low-rated topics, we suggest that its business owner can make overall improvements in these aspects.

5 Conclusion

Our analysis focused on Yelp datasets Businesses and Reviews, and used Trips and Census dataset as supplements. Our analysis on Businesses dataset produced some realistic recommendations, but the small dataset limited their statistical potency. In contrast, our LDA model on a large dataset produced significant recommendations, but they could be too abstract to interpret. After reviewing our project, we think a better approach would be to have a broader topic so that more information from Trips and Census dataset could be incorporated. At the same time, we should filter reviews more so that we would only deal with typical reviews (e.g. extreme star rating and extreme sentiment score) . If we balance our datasets this way, we would get more balanced recommendations.

Reference & Contribution

[1] Kang, H. (n.d.). Yelp Documentation. Yelp dataset. https://www.yelp.com/dataset/documentation/main
[2]Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland. (2023, November 28). Trips by distance: Tyler Data & Insights. Bureau of Transportation Statistics. https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv
[3] Bureau, U. C. (2023, November 23). Census.gov. https://www.census.gov/

Table 1: Member's Contribution

Contributions	**Ruofeng Tang**	**Ziming Li**	**Bingyan Liang**
Presentation 1	Produced page 1-8	Produced page 9-12	Produced page 13-19
Presentation 2	Reviewed all pages	Reviewed all pages	Produced all pages
Summary	Produced first draft	Produced LDA analysis part	Reviewed all parts
Code	Responsible for Analysis on the Business, Trips and Census datasets	Responsible for LDA Analysis and the Reviews dataset	Reviewed and organized all codes in Github, cleaned Data
Shiny app	Contribute to shinyapp	Contribute to shinyapp	Reviewed all codes