Yelp Data Analysis - Mexican Restaurants in Hillsborough, FL

Group 1: Ruofeng Tang, Ziming Li, Bingyan Liang

Data Preprocess

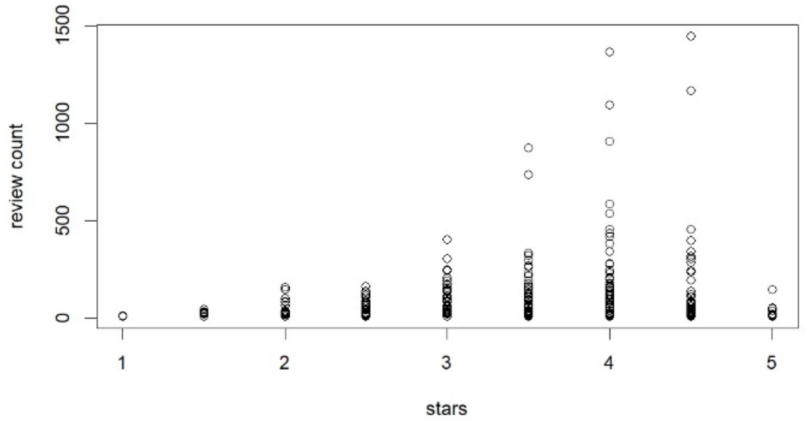
Dataset Sources: Yelp, U.S. Bureau of Transportation Statistics, U.S. Census Bureau

Data Preprocessing Steps:

- Extraction of relevant Mexican restaurants in Hillsborough County, FL, from Yelp
- Flattening of the 'attributes' column
- Reverse geocoding for obtaining zip codes
- Merging Businesses and Reviews datasets
- Filtering and processing U.S. Bureau of Transportation Statistics data
- Manual recording of demographic information from the U.S. Census Bureau
- Merging demographic data with the Businesses dataset

Analysis on Business Dataset: Plot analysis

More restaurants had high star ratings and high review counts.



Star Rating vs. Review Count Plot

Analysis on Business Dataset: Two-Sample t-tests on Restaurant Attributes

Result:

- 11 attributes with over 100 records: Bonferroni correction for alpha=0.05 gives adjusted alpha=0.0045
- Two significant attributes: **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 suggests that the removal of a service would increase star rating

Possible Reasons:

- Extra services may cause extra problems to complain about in reviews
- Small dataset contains only 329 rows, causing the results to be not as significant as later analyses.

Analysis on Business Dataset: Influence of Latino population

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.

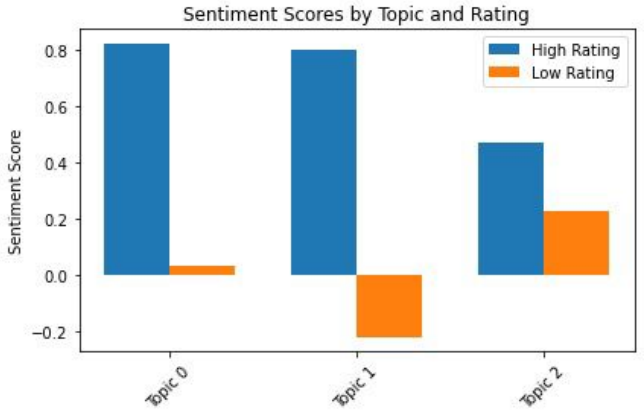
Our linear regression did not produce interpretable results, and no correlations were detected. However, we identified zip code regions 33603, 33604, 33614 in Tampa City, 33570 in Sun City as the best regions for Mexican restaurants.

The average star rating for each region is above 4, and there are a significant number of restaurants in each of them.

Analysis on Reviews dataset: Sentiment analysis and LDA

Highly rated topics(1-gram): positive aspects, such as “delicious food” or love for the restaurants in general. High average sentiment scores.
Topic 0 (**quality of food**): emotion score is 0.836.

Low rated topics(1-gram): “**Chipotle**’ , delays and order issues are always mentioned in the negative reviews. Sentiment scores are low or negative.
Topic 1 (**Ordering experience and customer service**): emotion score is -0.219.



Analysis on Reviews dataset: Sentiment analysis and LDA

2-gram and 3-gram topic analysis:

Similarly, the highly rated topics focus on the quality of the food, the low-rated topics still mentioned “Chipotle” and operational issues (order, lines, services).

Highly rated topics generally have higher sentiment scores, while the sentiment distribution of low-rated topics is uneven.

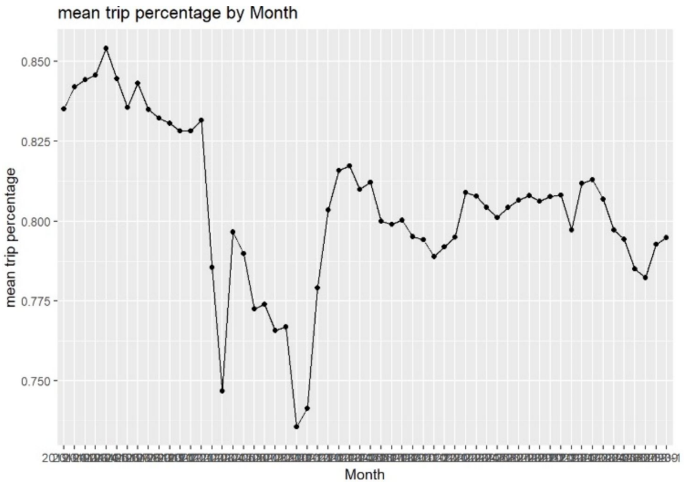
This may indicate that there are also some positive comments among some low-rated comments, but only for specific aspects, such as food-related.

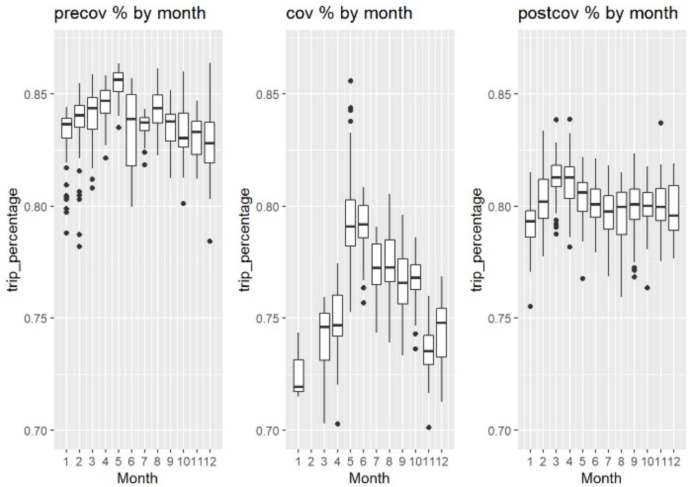
Analysis on Reviews dataset: Statistical significance and model fit evaluation

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.

Analysis on Trips by Distance Dataset : Impact of COVID

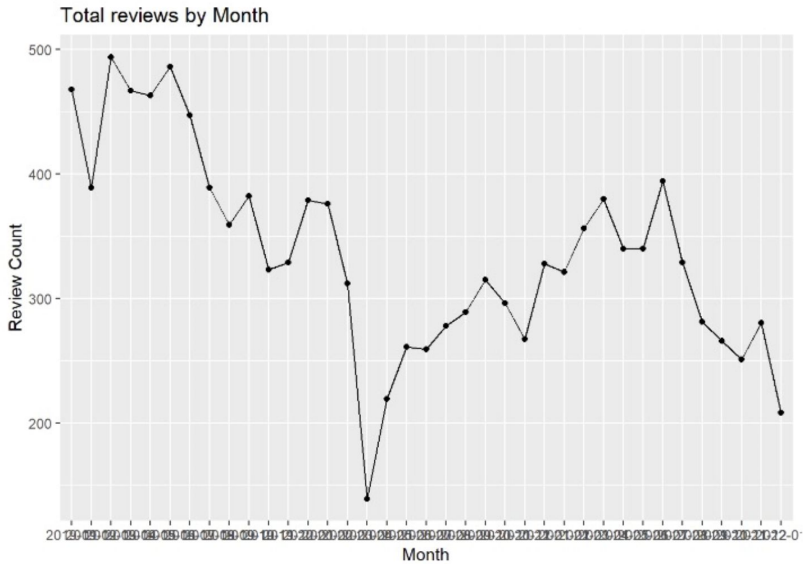
The comparison graph displays the percentage of people traveling per month before, during and after Covid, a clear drop occurred during Covid, and it did not recover after the pandemic.





Analysis on Trips by Distance Dataset : Impact of COVID

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.



Limitations

- 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.

Conclusion: Recommendations for Merchants

- **Geographical Recommendations**: Zip codes 33603, 33604, 33614 in Tampa City, or 33570 in Sun City
- **Restaurant Attributes**: No TV, cautious with delivery service
- **Quality Control:** considerations for sour cream
- **Improvements:** ordering process, staff training, especially for managers
- **Topics related to "Chipotle":** Overall improvements are needed