Using Yelp Data to Assess COVID-19's Impact on US Restaurants

Team 018: Yanlin Chen (ychen3402), Fangzhou He (fhe48),

Xinyue Li (xli901), Jianan Li (jli3059), Yaozhen Li (yli3307)

Abstract

The COVID-19 pandemic started in early 2020 and has caused huge damage to nearly all industries over the US economy, especially over small businesses that provide face-to-face service. This project explored the impact of COVID-19 on US restaurant industry with restaurant information from Yelp.

Introduction & Motivation

The outbreak of COVID-19 has shaken the world in an unprecedented way. The impact of the coronavirus pandemic on the global economy is inestimable [1]. In US, with the lockdown spreading the whole country, tens of thousands of small businesses must close their stores and shut down their businesses [2]. Based on the survey conducted between March 28 and April 4, 2020, 43% of small business were temporarily closed [3]. During this time, COVID-19 pandemic brings out the high-demanding of stay-at-home orders, which intensified the hit on US restaurant industry [4]. However, not much available resources provide a visualization feature to the public users in a friendly way [HQ2]. In this research, we are going to predict the likelihood of restaurant's closing. geographically in US, so that both the restaurant owner and local governor could plan for the next business model and financial strategies [HQ1&4].

Problem Definition

To assess the impact of COVID-19 on the restaurant industry, the objective is to analyze restaurant and review data from Yelp and gain insights that is beneficial to restaurant owners, employees and customers, as well as governments and local authorities. This project will create a Tableau dashboard on public repository for users to foresee an area's closing ratio and adjust their business model based on local favorite service [HQ1]. The estimation is based on sentiment score through customers' review attitudes. The result will generate a list of area they should improve on to lower their risk of closing [HQ4&5].

Proposed Method

Intuition

Based on our literature review, current studies mainly utilize two forms of data: tabular data from government and agencies (e.g., demographic information) and survey data (e.g., assessment of restaurant owners' perspectives on the pandemic development outlook [3]). Most of the studies are presented with an aggregated on the national level. This project aims to produce an interactive tool that provide analysis of restaurant closures on an adjustable level from state to zip code, which allows users (especially business owners) to gain insights more pertinent to their areas of interest (e.g., their business location).

Secondly, the tabular and survey data sources indeed provide a general understanding of the COVID-19's impact on restaurant's policies and economic performance, yet they are still incomprehensive in that they lack the perspective from customers, which is, in most cases, hard to obtain and measure for the researchers, but very crucial for business owners and policy makers to make decisions. Besides the evaluation and analysis of restaurant closures from restaurant information, this project aims to explore insights from customer's perspective by analyzing customer ratings on restaurant and conducting sentiment analysis on customer reviews. Due to its unstructured characteristics, the customer review data will provide unique insights that are currently lacking in the literature we reviewed.

Data

We obtained over 200k records of small businesses' details and special services (e.g., delivery/takeout enabled) offered during COVID-19 and 8M customer reviews from Yelp's public sources. The business detail data and the customer review data gathered are covered till 2/21/2020; the COVID-19 dataset

provides a snapshot on 6/10/2020. After cleaning and merging, we gathered details and services for about 50k US restaurants. A limitation of the data is in its uneven distribution of the number of restaurants across states, and thus we selected NV, AZ, WS, IL, OH, PA, NC, SC to represent the national trend due to their more significant sample size. We believe scale proportional selection here is representative according to Leskovec's data mining [6]. Through exploratory data analysis, we found that 38.3% of the restaurants are closed on 6/10/2020, which supports the national survey results (43%) [2] and implies that our states sample is an accurate description of the national condition.

To learn the most recent situation of these restaurants, we collected the information for the 50k restaurants from Yelp's API on 4/1/2021 to gather their updated information on restaurant closure. In addition, we gathered Yelp's updated business and review dataset till 1/28/2021 to have a better grasp of customers' review before and after the peak of the pandemic (we assume 1/28/2021 to be in the recovery phase of the pandemic). Finally, due to the large size of the review data, we used 5% as sample to select, design and validate our algorithm and then executed the algorithms of choice with about 20% of the review data.

Algorithm

Definition of Close Ratio is the proportion of restaurants that are:

1) opened on 2/21/2020 (representing before COVID-19)

2) closed on 4/1/2021 (representing after COVID-19)

Out of the total number of opened restaurants on 2/21/2020. We calculate close ratios on various dimension like price range, restaurant rating and location to assess the longer-term impact of COVID-19 on restaurant.

To setup the sentiment analysis, we use NLTK features and classifiers to detects polarity [7]; a positive or negative opinion; within customer reviews [8][9]. According to [10], VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities. The sentiment score of a text can be obtained by summing up the intensity of each word in the text [11]. The score indicates customer's attitudes in the range of [-1,1]. Negative score indicates that customer's negative attitudes in reviews, vice versa. Thus, we can measure customer's attitudes by analyzing their reviews [12].

Also, we use Latent Dirichlet Allocation (LDA) [13] to generate topics for each category, so that we will know what customers are concerning about under different circumstances. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions [13]. Each word in the distribution of words in a topic is similar to a side of the dice, and we have Dirichlet parameter to control if all the words have same probability in a topic or will that topic have an extreme bias towards some words. Due to the significant amount of data and API's limitation, we created 6 accounts for each ran the model through AWS and have spent $20 by far for 20% of the sample size [HQ7]. We would estimate 30 more hours for running through the entire data size [HQ8].

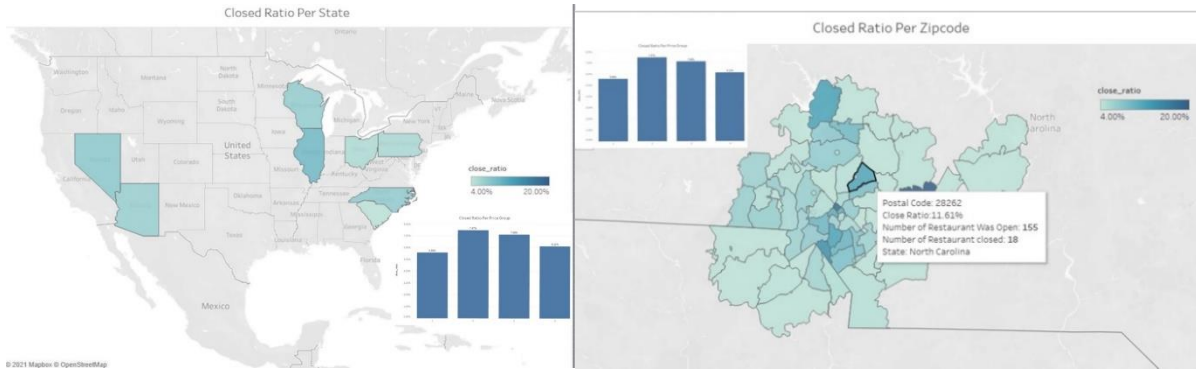
Based on the LDA topic model, we clustered reviews into 20 clusters. The visualization in [Appendix 1] serves as a very good representation of the distribution of a certain topic. The edges of the apex points indicate that the probability of some word to belong to a topic reduces to null [5].

Interactive Visualization Tool

To assess COVID-19's early impact on restaurants, we have a map visualization of the closing ratio for state field during pandemic. For each mouse over region, detailed close ratio information will be display in tooltips and closed group shows in bar, as showing in the left figure below. Users could select a detail zip code area for more specific insights, which will generate a detail local bar close group and local information in tooltips, as showing the right graph below. (Details in appendix and in tableau public repository here)

Utilizing the sentiment analysis and topic model algorithms, we are able to derive sentiment score and summarize the unstructured review data into topic labels. Therefore, we produced various visualizations

and insight-generating tools to excavate the value of customer reviews. (Details later)



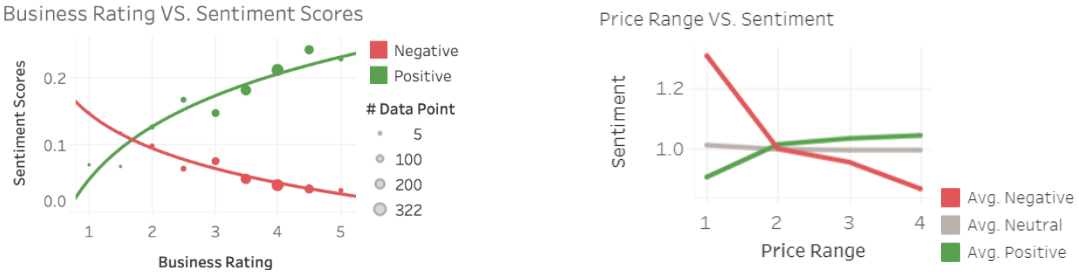
Map visualization of the closing ratio for state field during pandemic

Experiment and Results

The analysis of business information and restaurant's covid policy is mainly used to produce the interactive tool. The experiment in this project is mainly related to the review dataset through the usage of sentiment analysis and topic model as mentioned in the Algorithm section. The review dataset was split into two parts: before / after covid according to the review timestamp's relationship with 3/1/2020.

We used NLP and LDA models to evaluate and foresee aspects that the restaurant should improve to avoid closure. Besides, we flipped the use of cluster to anticipate potential closing risks area based on the sample feature while still accounting local trend. We confirmed the validity of our approach to use LDA on yelp data to generate operational insights from Luo & Xu [14].

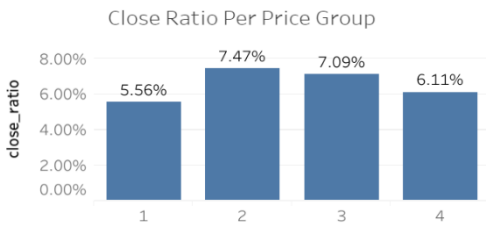
Evaluation of Results Validity



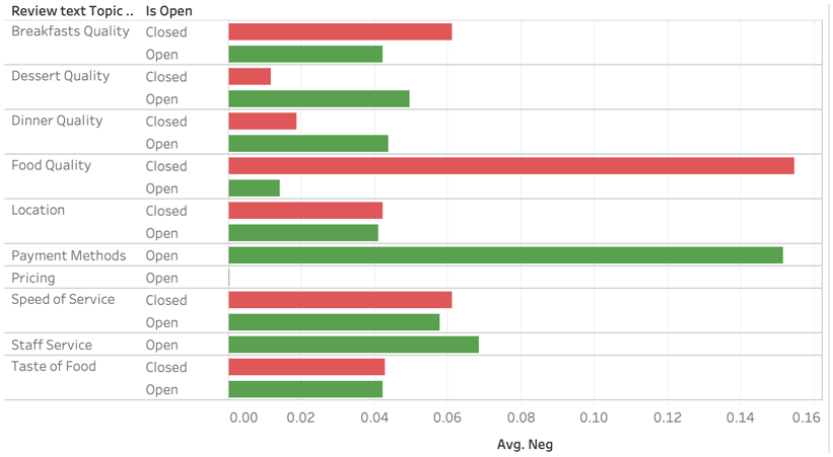
From the left chart, Business rating is positively related to positive scores and negatively related to negative scores. This chart verifies that sentiment score is a good indicator on customers opinions. From the right chart, higher priced restaurants tend to have more positive reviews, while lower priced restaurants tend to have more negative reviews. The results from these two graphs comply with common senses and therefore validate our usage of sentiment analysis on customer reviews to derive insights.

To further validate our experiment and approach, we collected 22 results from our survey. On the question "Which characteristic of restaurant do you think make them more prone to close due to COVID-19?": 41% of the respondents chose the price level; 27% chose location; 9% chose food quality and 23% chose restaurant type. These results correspond with our visualization analysis, which will be discussed later. On the question "Which aspect about restaurants do you care more about during COVID?" 50% of the respondents chose the delivery/pickup options; 18% chose the menu/food quality; 32% chose the price level. These results match the topic model labels analysis result, which will be discussed later.

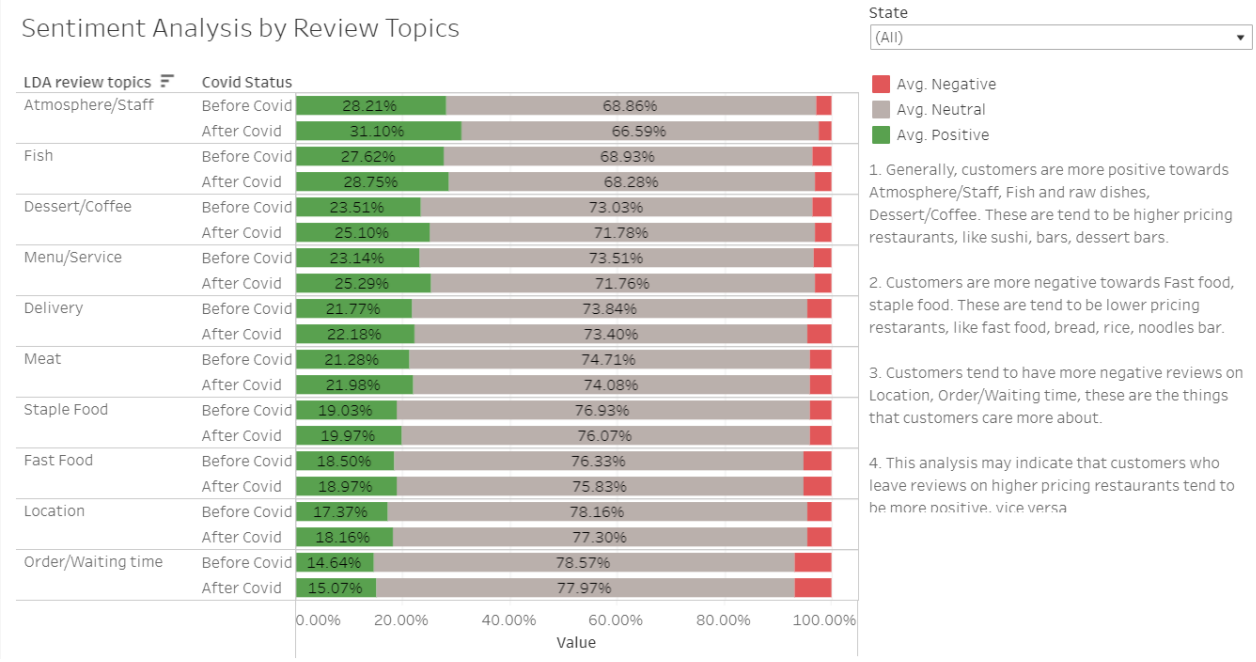
Experiment Observations



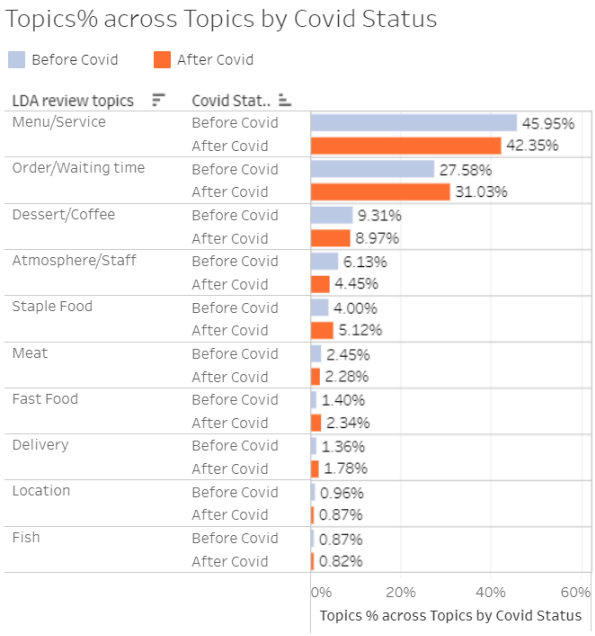
Firstly, Yelp use symbol "$" to give an estimate of restaurants' price levels: the higher priced restaurants have more "$". According to our literature review, restaurants priced at $ or $$ are 20% more likely to close during the pandemic, while restaurants priced at $$$ or $$$$ are 20%-100% less likely to close [15]. Referring to our geographic analysis, the result shows that the restaurants with $$$ and $$$$ are less likely to close than those with $$, which corresponds well to the literature review results.



Secondly, utilizing the COVID-19 dataset, we assessed aspects in customer reviews by restaurant closure condition on 6/10/2020, which was the peak of the pandemic. Based on the above table, restaurants that remained opened at the pandemic's peak have fewer negative comments on food quality, which implies that food quality is a crucial element for those restaurants that can get through the difficult time.



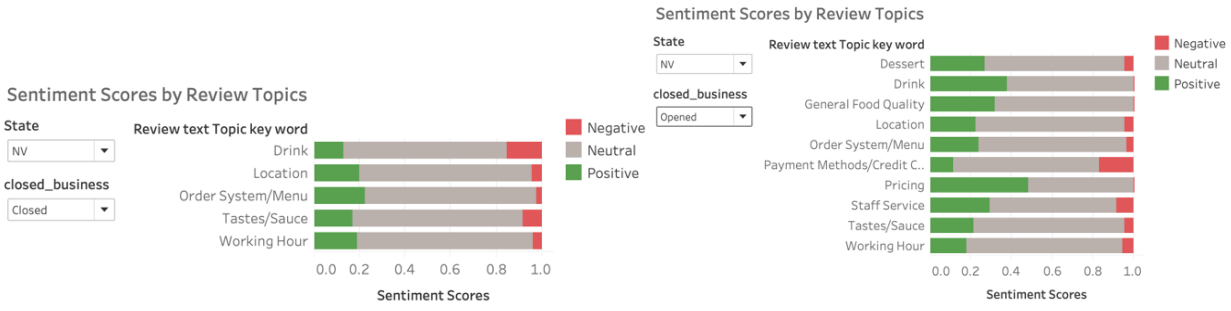
Thirdly, the observations from all the states are listed on the right side of the above visualization. This is also an interactive tool given the drop-down box to select different states, which enables restaurant owners to focus on a particular state of interest.



Finally, the above table is a summary of the review counts by topic by COVID-19 status (before or after Mar.1st 2020), in terms of percentage of the total review counts by COVID-19 status. The following observations can be drawn from the table:

1. Delivery is one of the most common way for restaurants to engage with customers in times of closure of in-door dining. Although "Delivery" only accounts for a small proportion of all reviews, there is a large increase (30.8%) of review counts from before to after Covid. This shows that customers tend to attach a larger importance to Delivery in restaurants' service after the pandemic.
2. The review counts of "Atmosphere/Staff" has decreased significantly (27.4%) after the pandemic, alongside with a minor decrease in the counts of "Location". This is also very reasonable since many customers choose to order by delivery and pickups, instead of dining-in, and thus the atmosphere and staff has a much less influence on customers' experiences. Similarly, customers care less about the location of the restaurant because a lot of them transfer to delivery during the pandemic and the physical location becomes less important.
3. There is a decrease in the mentioning of Menu/Service, which aligns with the results that show customers talking less about Dessert/Coffee and Fish, but more about Staple Food (increase by 28%) and Fast Food (increase by 67%). This indicates customers shift of consumption towards more necessary food and less high-end dining (Dessert and Fish) after the pandemic.

Discussion & Extension of Experiment and Visualization Tool



According to above tables, restaurant owners can directly see the customers' attitude towards each metrics. For example, when we compare closed restaurants with opened restaurants in Nevada, we can see that customers hold a strong negative attitude on “Drink” towards closed restaurants. For opened restaurants, they have a lot of positive reviews on “Drink”, with a small number of negative ones. Thus, we know that Beverages and Drink is an important metric for merchants as customers in this area really care about it. If a restaurant wants to survive in the market, it shall pay more attention to improve their drinks.

The above analysis shows an example of how restaurant owners can utilize customers’ review topics to generate insights on business strategies. In fact, our type of experiment and visualization approach can be applicable to other use cases, especially where a before-after effect analysis for a public event is needed. For example, since the Yelp dataset contains a much wider range of businesses beyond restaurants, an analysis can be done on the effect of social justice movements on small businesses that has a special focus on minorities. The scalability of our approach

From another perspective, our approach may act as an economical substitute and an insightful supplement to surveys, which are widely used in most of our literature reviews. Due to its higher cost to conduct and unpredictive participation rate, retrieved surveys are usually limited in its quantities, and thus only provide a small-size sample of the population, which may also include sampling and selection bias. Moreover, surveys usually limit their questions to multiple choices in order to attain a better chance to be completed and retrieved, and thus the respondents’ true intention might be limited by the available choices. However, in our project, we collected about 8M reviews for 50k US restaurants at almost no cost, and the size of our dataset would ensure sufficient sample size for conclusions we draw. More importantly, the nature of our topic model makes it possible to analyze on the true intents of the customers and thus capturing the benefit of unstructured data.

Conclusion

This project investigated the impact of the COVID-19 pandemic on US restaurant industry from the lens of restaurants and customer review information from Yelp. Through literature review, we gained insights on the general effect of the pandemic on restaurants and had a better understanding of currently used algorithms and research approaches. With the design, implementation and testing of our interactive visualization tools, as well as the sentiment analysis and topic model algorithms, we produced tools and insights that can help different stakeholders to acquire information.

Specifically, all members contributed with similar amount of effort, but with the emphases as follows:

Yanlin Chen: NLP & LDA topic models & Sentiment and Topic Visualization Tool

Fangzhou He: Data filtering and cleaning & Final Reports drafting

Yaozhen Li: Exploratory Data Analysis & API Scraping & Geographic Visualization Tool

Xinyue Li: Proposal and Progress Report Drafting & Poster Design

Jianan Li: Proposal and Progress Report Drafting & Poster Design

Throughout this project, we have learned a lot about conducting data analysis on large, structured or unstructured datasets, as well as creating meaningful and user-friendly visualization tools. We sincerely hope that we have added value to the society by creating the tools and analysis for the COVID-19 pandemic.

References

[HQ1 - 9] Heilmeier Questions. Available at: https://en.wikipedia.org/wiki/George_H._Heilmeier

[1] Debata, B., Patnaik, P., & Mishra, A. (2020). *COVID -19 pandemic! It's impact on people, economy, and environment*. Journal of Public Affairs, 1–5. https://doi.org/10.1002/pa.2372

[2] Kim, J., Kim, J., & Wang, Y. (2021, January). *Uncertainty risks and Strategic reaction of Restaurant firms amid COVID-19: Evidence from China*. Retrieved March 13, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7666809/

[3] Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., & Stanton, C. (2020). *The impact of covid-19 on small business outcomes and expectations.* Proceedings of the National Academy of Sciences, 117(30), 17656-17666. doi:10.1073/pnas.2006991117

[4] Yang, Y., Liu, H., & Chen, X. (2020). *COVID-19 and restaurant demand: Early effects of the pandemic and stay-at-home orders*. Emerald Insight, 32(231), 3809-3834. doi:10.1108/IJCHM-06-2020-0504

[5] Humphries, J. E., Neilson, C., & Ulyssea, G. (2020, April 26). The evolving impacts of COVID-19 on small businesses since the CARES Act. Cowles Foundation Discussion Paper.

[6] Leskovec, J., Rajaraman, A., Ullman, J (2014). *Mining of massive datasets: Chapter 9*

[7] Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine Learning Algorithms: A case study of Yelp. Sustainability, 11(19), 5254. doi:10.3390/su11195254

[8] Tardin, M. G., & Perin, M. G. (2020, October 19). *The Impact of COVID-19 on the Brazilian Food Service Industry: Topic Modelling of Online Reviews*. Retrieved from http://bibliotecadigital.fgv.br/ocs/index.php/clav/clav2020/paper/viewPaper/7588

[9] Pei-Ju, L.T, Hsiang C, Wen-Chang F and Szu-Ling C (June 2017). *Using Big Data and Text Analytics to Understand How Customer Experiences Posted on Yelp.com Impact the Hospitality Industry*. Page 107-130, Vol. 13, No. 2

[10] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[11] Deng, L., Liu, Y., Ullman, J (2014). Deep Learning in Natural Language Processing*: Chapter 8*

[12] Yao, Yao; Angelov, Ivelin; Rasmus-Vorrath, Jack; Lee, Mooyoung; and Engels, Daniel W. (2018) *"Yelp's Review Filtering Algorithm,"* SMU Data Science Review: Vol. 1: No. 3, Article 3

[13] D. Blei, L. Carin and D. Dunson, "Probabilistic Topic Models," in IEEE Signal Processing Magazine, vol. 27, no. 6, pp. 55-65, Nov. 2010, doi: 10.1109/MSP.2010.938079. https://ieeexplore.ieee.org/abstract/document/5563111

[14] Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine Learning Algorithms: A case study of Yelp. Sustainability,11(19), 5254. doi:10.3390/su11195254

[15] Kramer, S. (2020, Dec). *Analyzing COVID-19 Restaurant Closures with Yelp Data*. https://medium.com/13-fund/analyzing-covid-19-restaurant-closures-with-yelp-data-f9116c7d563a

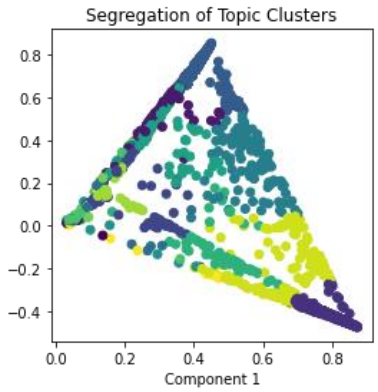
[16] Fairlie, R. W. (2020). *The impact of Covid-19 on small business owners evidence of early-stage losses from the April 2020 current population survey.* Cambridge, MA: National Bureau of Economic Research.

Appendix

Appendix 1

https://public.tableau.com/profile/yaozhen1140#!/vizhome/Team018AssessCOVID-19ImpactonUSRestaurants_ThroughYelpData/Story1?publish=yes

Appendix 2



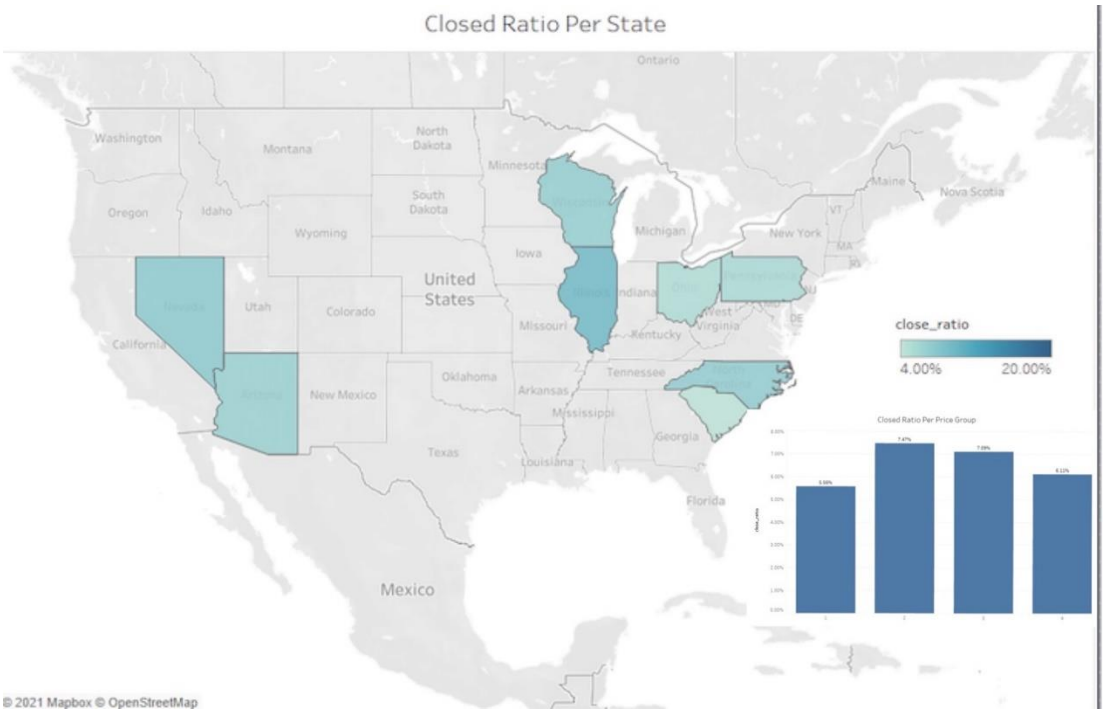
Appendix 3: Survey to consumers

Number of participants: 22

Question 1: Which characteristic of restaurant do you think make them more prone to close due to COVID-19?
A: price level B: location C: food quality D: restaurant type

Question 2: Which aspect about restaurants do you care more about during COVID?
A: Delivery/pickup options B: menu/food quality C: price

Appendix 4: State close rate map



Appendix 5: User's selection of zip code