Restaurants Performance During COVID-19: A Yelp Analysis

University of Toronto JSC370 Final Project

Shih-Ting (Cindy) Huang

04/21/2022

Contents

Introduction

After the emergence of COVID-19 in early 2020, undoubtedly many businesses were forced to close temporarily or permanently. Throughout this time, restaurants in particular have undergone various modifications, such as disabling dine-in, starting takeout, etc. As a result, the food quality, service quality, and overall dining experience might have changed - whether in a positive or negative way.

For this study, I utilized data from Yelp, one of the biggest platforms that displays information of businesses and enable users to search for, react to, and/or share their opinions of such businesses. Their data come from a total of 6,990,280 reviews and 150,346 businesses from 11 metropolitan areas - Montreal, Calgary, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland. Only reviews that are recommended by Yelp are included. For privacy reasons, content of the datasets will not be disclosed.

Research Question

Having established the research's significance, I propose the question: How has COVID-19 affected restaurant performance and impression? Specifically, do restaurants with services that better comply with COVID restrictions necessarily have better ratings?

For the purpose of this research, we define the start of the pandemic as January 9th, 2020, which is when WHO reported globally that there has been a new coronavirus identified.

Methods

Data Access

The data were retrieved directly from Yelp's Open Dataset in a `.tar` compressed file. Extracting the respective json files and converting each to csv using Python, we selected two relevant datasets - `business` and `review`, which provides details on a business and a review respectively. Since we are observing the impact of COVID-19, we decided to only take reviews from the most recent five years (starting at 2017/01/01) and limit data size.

First, we take a look at what variables are present in each dataset. For `business`, we have:

Table 1: Yelp Business Variable Definitions, obtained and modified from Yelp Dataset Documentation

Variable Name	Definition
business_id	22 character unique string business ID (str)
name	Business's name (str)
address	Full address of the business (str)
city	City (str)
state	2 character state code, if applicable (str)
postal_code	Postal code (str)
latitude	Latitude (dbl)
longitude	Longitude (dbl)
stars	Star rating from 0.0-5.0, rounded to half-stars (int)
review_count	Number of reviews (int)
is_open	Binary indicator for if business is open (=1) or closed (=0) (int)

Variable Name	Definition
attributes	A list of features associated with the business and whether it is available (True) or unavailable (False) (dict[str][bool])
categories	Categories that the business belongs to (str)
hours	Business operating hours on a 24-hr clock (str[dict])

Next, we take a look at the `review` dataset:

Table 2: Yelp Review Variable Definitions, obtained and modified from Yelp Dataset Documentation

Variable Name	Definition
review_id	22 character unique review id (str)
user_id	22 character unique user id (str)
business_id	22 character business id, maps to those in business dataset (str)
stars	Star rating (int)
date	Date formatted YYYY-MM-DD (str)
text	Review content (str)

A dataset for review ID and content (`rev_text`) was retrieved separately due to technical limitations.

Data Wrangling

Filtering Observations

Since we are only focusing on restaurant performance, we eliminated non-restaurant business in the `business` dataset and their corresponding reviews, retaining only those that are under the categories "Restaurants" or "Food". The list of categories that Yelp offers can be found here.

Originally, we have 150346 businesses and 3838105 reviews, and after filter we reduced to 21057 food businesses with 946469 reviews total.

Ensure data validity

Checking potential problems with the data types, we verify that there are 20 problems total, all of which arise from variable `postal_code` with the value `""`.

Checking for NAs

As previously determined, we notice that the datasets represent missing values as an empty string "", and below we showcase how many there are for each variable after converting"" to take on value `NA`:

Table 3: Summary of NA Proportion for All Variables (Yelp Business)

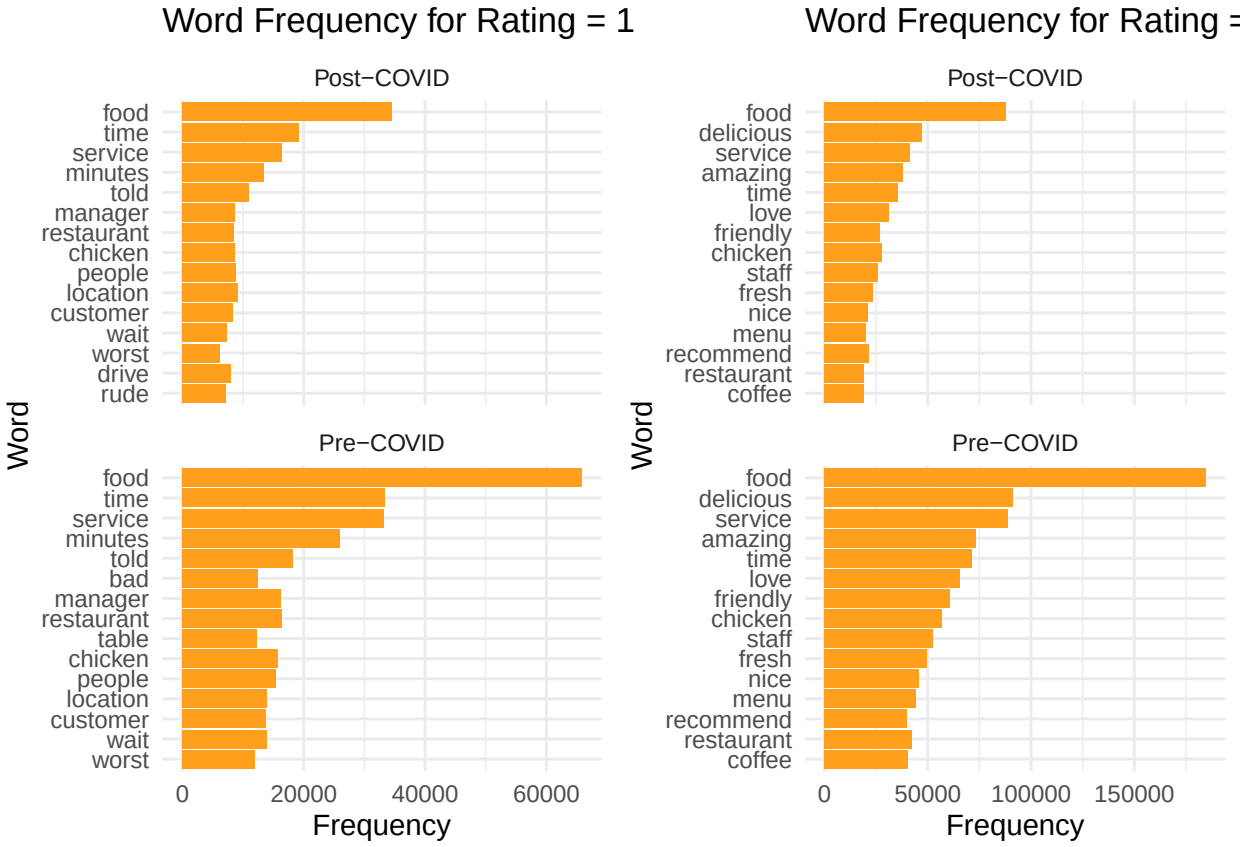
	Proportion
address	0.0190435
postal_code	0.0009498

	Proportion
attributes	0.0094980
hours	0.1005841

We notice that there are no missing values in `review`; however, in `business`, `NA`s are present in the variables `address`, `postal_code`, `attributes`, `categories`, and `hours`. Since these variables are unique to the restaurant and cannot be simply inferred from observations with non-missing values, we will keep them as NAs in the dataset as they may have particular association with other characteristics.

Natural Language Processing

For the reviews, we are interested in seeing how reviewers' opinion of the restaurants have changed overtime. In order to attain this, we decide to use sentiment analysis and determine the most common positive and negative words for each restaurant every year.This could potentially give us insights on features that are most associated with positive or negative reviews and hence the restaurant's performance.



For concise visual demonstration, I have selected the graphs for the two extreme ratings (refer to website for the other ones). Based on the most frequent words at every rating, we saw that topics related to the terms "food", "service", "time", "chicken", and "fries" are popular both before and after COVID. Some distinct terms related to bad ratings (< 3) are "table", "location", and "wait", whereas above average ratings ($>=3$) often mention the terms "fresh", "menu", "time", and "coffee". Based on these graphs, there doesn't seem to be a major difference in the most common terms between pre- and post-COVID reviews. Nonetheless, it suggests several business features we can consider for the model - ones related to location, waiting time, table service, and food category, specifically whether it is a coffee shop or not.

Feature Engineering

Based on the insights obtained from EDA, we generated additional features from existing ones that may be helpful for our data exploration and analysis.

89.5% of the businesses opened before COVID-19 whereas the remaining 10.5% have undetermined opening time. Such skew makes the variable likely unhelpful for our analysis.

We then extracted information from `attributes` and `hours`, which are dictionaries of restaurant features and open hours respectively. For indicator variables, we assigned the value 2 for unknown because the lack of such information could reveal insights about the restaurant performance as well. At this point, all of the variables we have added are:

- `avg_hours`: the daily average open hours (float)
- `takeout`: indicator for if restaurant does (=1) or does not offer (=0) takeout, or unknown (=2) (int)
- `delivery`: indicator for if restaurant does (=1) or does not offer (=0) delivery, or unknown (=2) (int)
- `takeout_deli`: categorical for if restaurant offers takeout, delivery, both, or neither (chr)
- `good_for_groups`: indicator for if restaurant is (=1) or is not (=0) good for groups, or unknown (=2) (int)
- `outdoors`: indicator for if restaurant does (=1) or does not offer (=0) outdoor seating, or unknown (=2) (int)
- `table_serv`: indicator for if the restaurant does (=1) or does not offer (=0) table service, or unknown (=2) (int)
- `drive_thru`: indicator for if the restaurant does (=1) or does not offer (=0) drive-through, or unknown (=2) (int)
- `is_cafe`: indicator for if the business is (=1) or is not (=0) a cafe/coffee shop (int)
- `price_range`: restaurant price range from 1-4 (int)

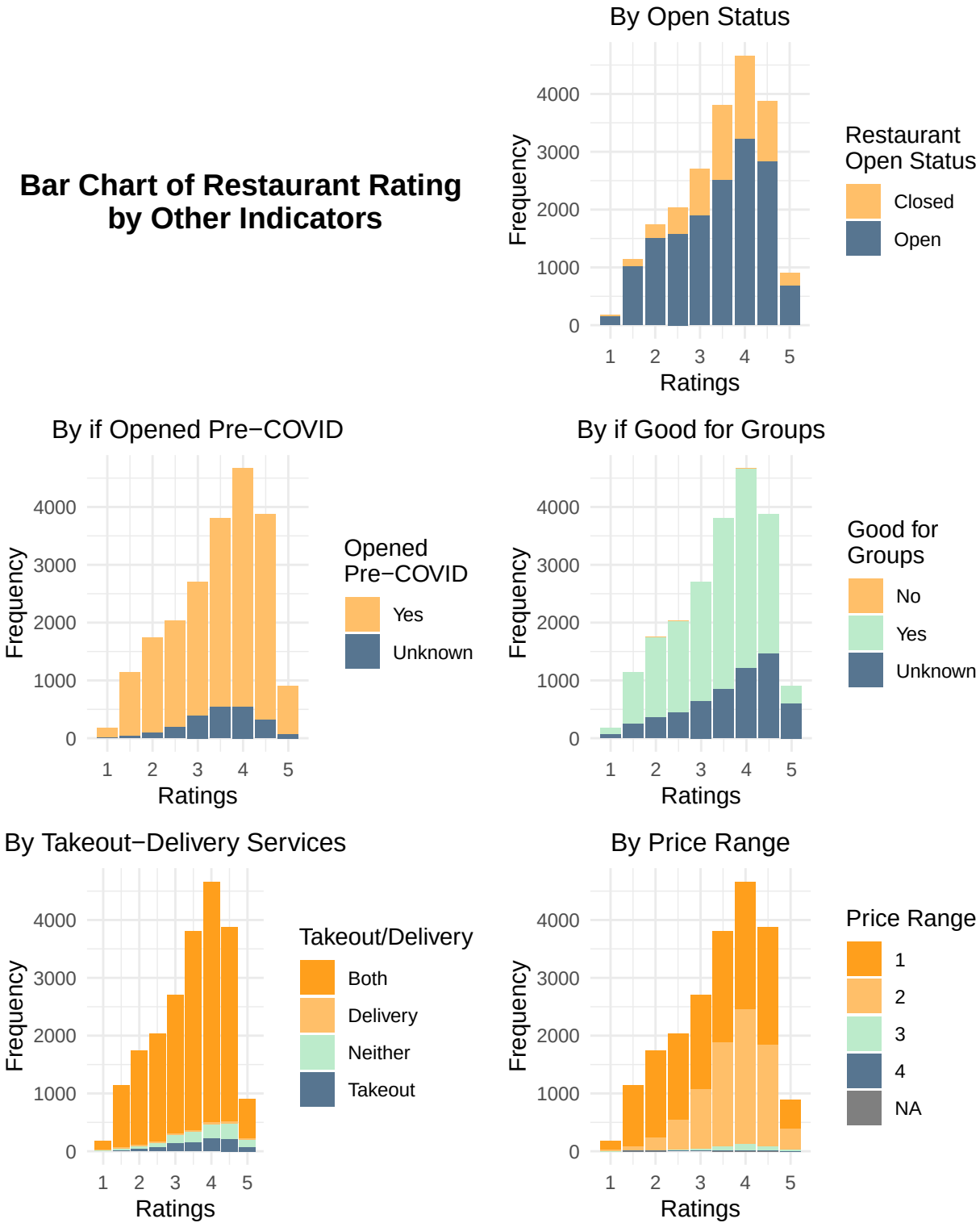
Tools used

Data wrangling were completed with `tidyverse` and `dplyr`. All figures were created with `ggplot2`, interactive visuals were created using `plotly`, and maps were created using `leaflet`. Tables were created with `kable` and `kableExtra`. Packages used for modelling & classification include `rpart`, `randomForest`, `xgboost`, and `caret`. `tidytext` was utilized for NLP.

Results

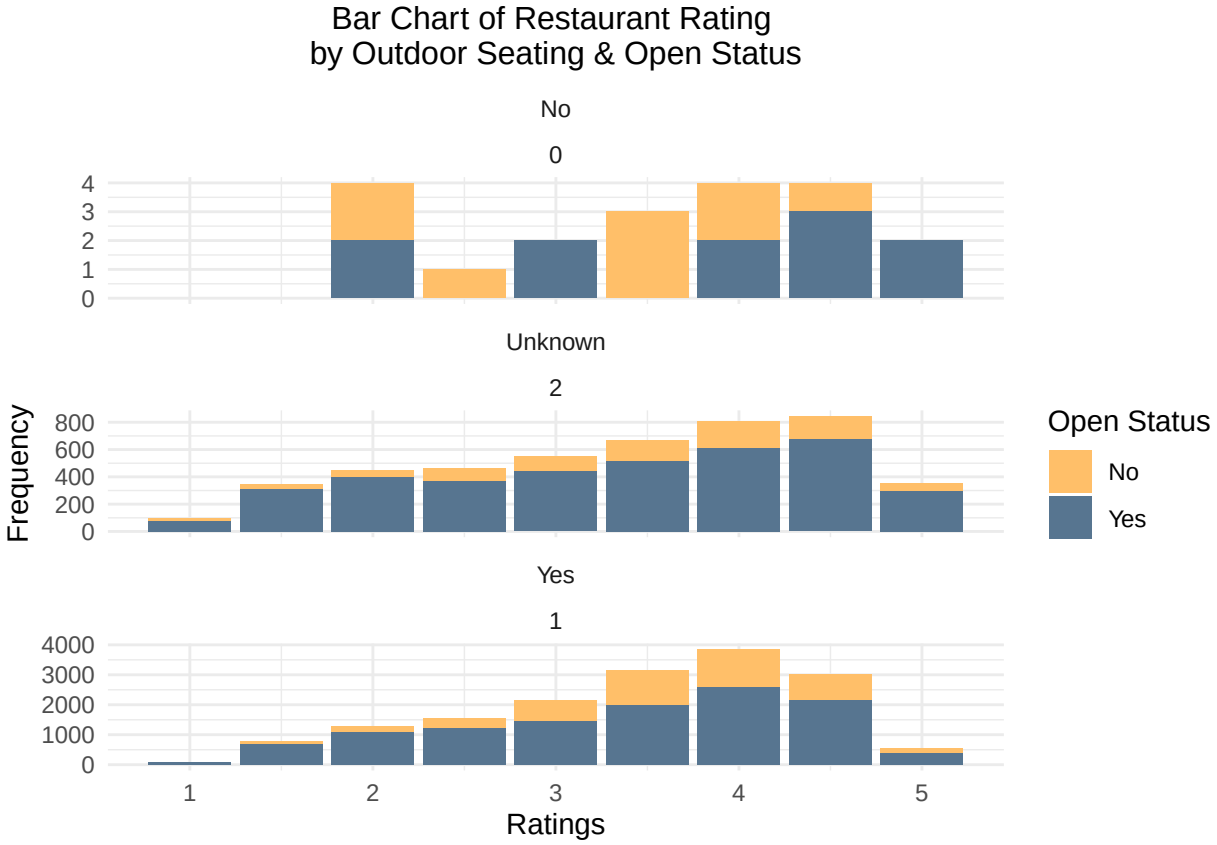
Summary Visuals

Multi-variable Relationships



It is visible that a majority of the restaurants in our dataset have 4 stars. Looking at the distribution of ratings stacked by various indicator variables, the most significant relationship we see is that a high proportion of restaurants with rating 4 are good for groups and offer takeout/delivery. The association with open status, surprisingly, reveals that a higher proportion among restaurants with ratings < 3 are still open. Ratings and price range do not have as clear of an effect on each other.

For some time during the past two years, COVID has discouraged/forbidden indoor dining in various places. Hence, this feature could potentially have an impact on how reviewers perceive the business.



Looking at ratings by outdoor seating availability and stacked by open status, we observe there are few restaurants that don't offer outdoor seats, out of which around 50% are still open with a rating of at least 3. On the other hand, those that for sure have this feature have a significantly higher proportion of open restaurants across all ratings.

Spatial Relationship

(Refer to website for interactive graph) Mapping out the locations of these restaurants, we note that there is no data for restaurant businesses in Toronto, Montreal, and Madison metropolitan areas as mentioned in the Yelp documentation. However, there are additions from Baton Rouge (Louisiana), Nashville (Tennessee), Tampa (Florida), Indianapolis (Indiana), Boise (Idaho), and Santa Maria (California). Although there is no clear trend between or within areas, an interesting observation is that within each area's cluster, more low-rated restaurants are located on the edge rather than in the middle, which supports NLP insights revealing complaints about location.

Time Relationship

(Refer to website for interactive graph) Since we are curious about pre versus post-COVID restaurant performance, we also examine how average review ratings have changed overtime, particularly for the business location (state) and whether they offer takeout/delivery or not. Here, we only included groups that averaged across more than 10 restaurants to ensure that we don't interpret trends based on insufficient data.

In terms of takeout/delivery services, there has been a steady rise in ratings among restaurants offering both, particularly after 2020/03. The trends for the other categories are less apparent. With regards to state, some have on average higher ratings than others. This reminds us that when modelling, location variables should be considered as random effects.

Modelling

Generalized Linear Mixed Model

First, we note that our response variable, `stars` (ratings), is a categorical variable taking on the values 0, 0.5, 1.0,..., 5.0. Moreover, business ratings within the same state can't be considered as independent from each other. There is more competition in cities with more restaurants, which can impact how people rate it relative to other options in the area. Under these conditions, we would have considered using a generalized linear mixed model with a multinomial response. However, the implementation is currently unavailable, so we instead use a binomial target for if ratings is <=3 or >3.

This model has an AIC score of 1.818821×10^4. We note that significant variables are: `review_count`, `is_open`, `avg_hours`, `price_range = 2`, `price_range = 3`, `takeout_deli = Delivery`, `is_cafe` based on their associated p-values from Wald Z-test:

	Estimate	P-values
Baseline odds	3336.57	0.325
Review Count	1.02	0.000
Is Open	0.55	0.000
Average Open Hours	0.89	0.000
Price Range = 2	2.17	0.000
Price Range = 3	1.90	0.001
Offers Delivery	0.63	0.022
Is a Cafe	1.48	0.000

With that in mind, we attempt to see if a simpler model including only those features yields a better AIC score.

The new model yields an AIC of 1.8534031×10^4, which is higher than the original model. Hence, we want to opt for the original model rather than the simpler one in this case.

Interpretation of Values

The estimates above represent the change in odds of the restaurant having a rating > 3 for a unit increase in the (continuous) variable or for belonging to a certain level of a categorical variable.

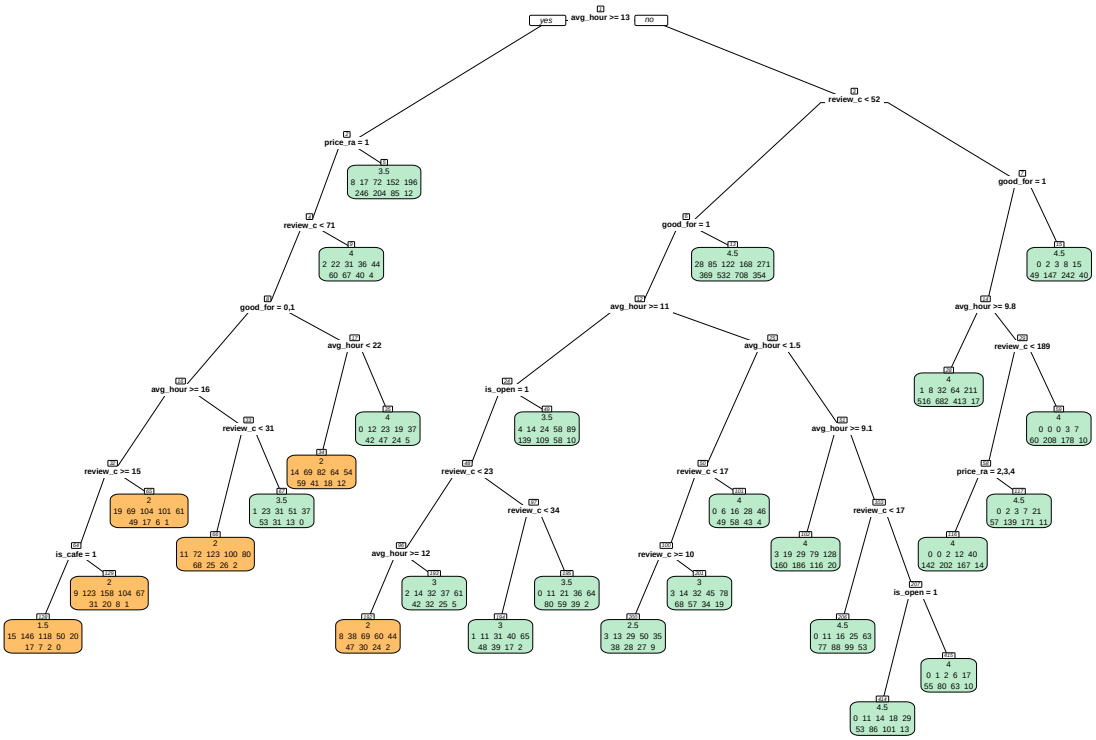
The estimates of the The odds of high ratings is 3336:1 when all other variables take their reference level (0). Looking at the variables with significant p-values, with one unit increase in reviews, we have a 2% increase in the odds of high ratings. Moreover, when the business has a price range 2, 3, or is a cafe, there is an increase in odds of high ratings by 117%, 90%, and 48% respectively. Average open hours and offering delivery, on the other hand, are associated with a 11% and 37% decrease in odds of high ratings.

Classification

Next, we try some machine learning models - Decision Tree, Random Forest, and XGBoost - to classify ratings with the significant features from GLMM and observe which features contribute to rating variation the most. With no restrictions on response type, we use the multi-class `stars` for more precision. Misclassification error instead of MSE will be used to compare model performance since we no longer have a binary outcome.

Decision Tree

Considering the amount of data we have, we decided to adjust the chosen hyperparameters accordingly. The class with the least number of observations is stars = 1.0 with 182 total businesses. As a compromise between model complexity and generalizability, we set the minimum cases before splitting and minimum leaf node to 200. Max tree depth has been set to 15.



Within each tree node, we have the predicted label as well as 9 values beneath showing the number of observations of each label (1,1.5,...,5) that has been assigned to that node. Colored in orange are the nodes with predicted value below 3.

Table 5: Decision Tree Variable Importance

Variable Name	Variable Importance
Average Open Hours	597.057622
Review Count	559.399264
Price Range	306.635790
Good for Groups	262.517987

Variable Name	Variable Importance
Is a Cafe	165.642854
Is Open	140.669197
Outdoor Dining	4.176022

Random Forest

No hyperparameters were specified for random forest since in building the trees, the model already attempts different value combinations.

Table 6: Random Forest Variable Importance

Variable Name	Variable Importance
Average Open Hours	774.04894
Review Count	677.67413
Price Range	163.83124
Good for Groups	113.70181
Is a Cafe	97.39083
Is Open	94.85844
Outdoor Dining	80.34547
Offers Takeout/Delivery	77.41643

XGBoost

For XGBoost, hyperparameters such as max depth, number of estimators, and learning rate are tuned for optimal performance. The chosen final model has the parameters: `nrounds`=50, `max_depth`=10, `eta`=0.3.

Table 7: XGBoost Variable Importance

Variable Name	Variable Importance
Average Open Hours	0.4759787
Review Count	0.2529025
Price Range = 2	0.0838237
Good for Groups = Unknown	0.0559289
Good for Groups = Yes	0.0474778
Is a Cafe	0.0354864
Is Open	0.0323440

Comparing the models' performance, we see that in general, the models aren't able to classify the ratings as well as expected. Out of the three options, random forest yielded slightly better results than the other two.

Table 8: Test set misclassification error on restaurant ratings

Model	Misclassification Error
Decision Tree	0.7149414
Random Forest	0.7062362
XGBoost	0.7200317

Due to this high misclassification rate, we take the variable importance outputs with a grain of salt and opt for results from the generalized linear mixed model, which gave us a clearer idea of response-feature relationship with the coefficient estimates.

Conclusion

From NLP, we found out that potentially, attributes related to location, waiting time, and table service correlate with different levels of ratings. There was also the prominence of the term "coffee" in positive reviews, which was later also included in the model. In the generalized linear mixed model, we saw that many of the attributes we hypothesized to influence restaurant ratings during the pandemic (ex. outdoor dining options, accessible drive-through) did not have a significant correlation with the odds of receiving high ratings, but still cannot be removed from the GLM model due to higher AIC. On the other hand, number of reviews, average open hours, and price range contribute the most to classifying restaurant ratings, which is reasonable contextually. An interesting insight was that indeed, cafes generally have higher ratings, possibly implying that coffee shops are more likely to perform well even amidst the difficulties presented by COVID regulations. In conclusion, our initial hypothesis that businesses with attributes more favorable during COVID will have higher ratings was not supported by this study. Instead, review count, operation length, and price still continuously appear as contributors to higher ratings.

Limitations & Future Directions

There are several limitations to the current study. For one, Yelp certainly isn't representative of the entire population. It is possible that there are hidden shared traits among Yelp reviewers that lead them to behave in a particular way.Additionally, there isn't enough information on changes overtime. Restaurants might have implemented takeout/delivery after COVID-19, or perhaps it reopened after a year of closure. We have no access to the actual situation of all businesses and hence can't isolate each variable's effects. Lastly, Yelp ratings is only one of the measurements of restaurant performance. For example, the number of reviews and the percentage of positive reviews (among others) are also valid metrics for this research, and can be investigated in the future.