Madalyn Li
DATA 512
Fall 2021

Analysis on Covid-19 & the Wake County Restaurant Industry

Introduction

Since the start of the pandemic, there has been a huge shift in how we socialize and interact with one another. One subset of this change is how businesses or restaurants have implemented stricter measures to ensure the safety and well-being of their customers. More specifically, in response to the pandemic, many restaurants have since offered take-out and delivery options to provide customers with the option to support businesses without having to compromise their safety. In this project, I aim to explore if these adaptations have helped restaurants thrive in business during the pandemic because they accommodated for the transformation in social culture and norms by providing safer options for customers.

An article published by Forbes in July of 2020 pointed to the fact that the restaurant industry and consumers' needs have changed significantly since the start of the pandemic. They cite a study conducted by SafetyCulture and YouGov that "found that 52% of Americans would feel uncomfortable dining in a restaurant/bar throughout the next 3 months". Moreover, the article stated that "concepts with heavy on-premise operations that don't have the ability to shift quickly to carryout or delivery are in a particularly precarious position; perhaps even worse than March and April when the nationwide lockdowns first went into place". Furthermore, a survey conducted by the National Restaurant Association in September 2021, emphasized the major shift and changes that severely impacted the restaurant industry, specifically regarding profit margins and business operations compared to pre-pandemic numbers. In one survey, they pointed out that 85% of restaurant operators indicated that their current profit margins were lower than their pre-covid numbers.

With this information, I do believe it is important to address and validate how society's needs for the restaurant industry have changed since the start of the pandemic. From a practical perspective, it will be interesting to see how this social shift, which has since caused many restaurants to offer takeout and delivery options, has potentially improved the trust and comfort level for customers. On the other hand, I hope that the findings will show that restaurants who could not shift quickly to carryout or delivery options did not perform as well in comparison. Regarding the human centered aspect, I believe that the analysis and corresponding results will support human needs and benefits. By validating how society responded to the notion of delivery and takeout options during the pandemic, other restaurants who are still potentially hesitant to offer this option and are on the brink of closure, may see the benefits of doing

so and alter their initial views. This can help drive more revenue for the restaurant industry and provide more options for customers who prefer to eat out and support local businesses. Finally, these results can help create an overall safer public environment that encourages and informs others to consider the safety and well-being of the entire population.

As I will further expand on in the subsequent section, the topic of takeout or delivery has been a contested research subject. Since the start of the pandemic, there have been many studies trying to understand the impact on the current restaurant industry and how the future of the industry will shift as a result. I do believe that this research will help to solve a real-world problem by helping others understand how the restaurant industry, which makes up a large part of the U.S. economy, has been impacted due to the pandemic. If we can foresee any shifts to customers' preferences in the near future, we can help forecast how the industry will perform.

Background/Related Work

Other research on this topic has focused on the shift in societal preference to takeout or delivery, but does not entirely focus on how it relates to a restaurant's Yelp ratings or reviews. An article published by Yelp cites that 90% of consumers indicated that they plan to eat at home just as frequently, if not more, going forward due to convenience and comfort. Furthermore, the article states that the pandemic has shifted how consumers view the dining experience and noted that nearly 70% of American adults said that they were more likely to order takeout now than they would pre-pandemic. On the other hand, in an article published by the New York Times in July 2021, they highlighted the continued struggles of restaurants needing to accommodate for the significant increase in demand in delivery options. They pointed out that it was difficult for restaurants to balance profit margins with additional delivery and advertising fees for delivery apps such as DoorDash and Uber Eats.

From these articles, it seems that the industry is trending towards an increase of demand for takeout and delivery, however, restaurants seem to be struggling to keep and and support this demand while trying to make their business profitable. I was interested to know how society's demand for takeout and delivery shifted as a result of the pandemic and how this positively or negatively impacted their perception of restaurants. Specifically, I wanted to measure and compare restaurants that offer takeout and delivery and those that did not to see their overall success during the pandemic measured by Yelp ratings and reviews. The reason why I chose Yelp as a measuring point for success is because it is such a prominent tool used by a large majority of Americans today to decide whether or not they want to eat at a restaurant. Typically, Yelp ratings are used as a simple indicator for a consumer to see and gauge how well others

liked a restaurant. Additionally, it can also be used as a tool for businesses to understand what improvements they can make to provide a better restaurant experience for customers.

My research question is: Have restaurants in Wake County, NC, who have adapted measures to handle the pandemic, specifically around offering takeout or delivery options, seen any change in their overall ratings and foot traffic compared to restaurants who did not choose to adapt? My prediction is that I will see a positive change in average ratings and foot traffic numbers for restaurants that did choose to adopt stricter measures during the pandemic. The logical reasoning behind my prediction is, since Wake County is a fairly progressive city (as shown by the results that nearly 72% of people would always wear a mask in public from the New York Time's Masking Survey Data), the data will reveal that most consumers in the area would feel uncomfortable dining out at restaurants. Thus, the businesses that opt for takeout and delivery options would see more foot traffic and higher ratings because of the consideration for customer preference.

Methodology

My overarching method and plan for this project was to gather the necessary data needed for analysis, cleanse the data, join the results, and calculate the measures needed for analysis to produce the final visualizations. The goals for the end results of this project were to produce two time series graphs for the average Yelp ratings and the total number of Yelp reviews plotted over time to see the monthly changes for these values between restaurants who do and do not offer takeout or delivery. In addition, I also planned to plot the monthly confirmed Covid-19 cases on a secondary axis to help users visualize the start of the pandemic and the waves and trends throughout its course. This would help support the understanding of how the pandemic truly affected the factors of restaurants' adaptations surrounding takeout or delivery.

The datasets I used for my analysis were obtained from three different sources: Restaurants in Wake County, Kaggle's Daily Covid-19 Data Repository, and the Yelp Fusion API. The restaurants in Wake County dataset contains a list of all restaurants within Wake County that can be found on Yelp. This dataset was used as a reference point to query business and review information from the Yelp Fusion API. The data queried from Yelp Fusion API provides valuable data points that were necessary for the final analysis. For instance, the business endpoint query provided details on business_id, review_count, rating, and transactions (which contains the labels for instances where a business offers 'pickup' or 'delivery'). Furthermore, the reviews endpoint query provided details on review_id, rating, and time created. One huge limitation to using the Yelp Fusion API was that it only returned the top three reviews for each

business ranked by Yelp's default sort order which takes into account recency, user votes and other factors. Later in this section, I will detail how I adjusted to this limitation by mitigating some bias and using the Chi-Squared test for data validation. The final dataset I used came from the Kaggle's Daily Covid-19 Data Repository, which was also used in Assignment 4: Common Analysis in this course. Regarding the ethical concerns for the use of Yelp data, I do not foresee any issues because Yelp is a public site that is generally used by many restaurants and consumers. All the data that I gathered on the restaurants are publicly available on the restaurant's own Yelp page for anyone to view at any time. In addition, the Wake County restaurants dataset is also publicly available and includes already public information whose purpose is to inform users on general business information. Lastly, the data for reviews from the Yelp Fusion API will not be tied to personal information of the users who left the review, as I only used it to count the total reviews and calculate the average ratings.

After gathering the data, the next step was to cleanse the data. The first step in this method was to mitigate bias from the reviews collected and to validate that the proportions of the variables were dependent. To mitigate bias, I removed any businesses from the reviews dataframe that did not contain reviews that were uploaded from timeframes before and during the pandemic. Additionally, I also removed any instances of businesses that only contained one review. This ensured that these data points would not have a significant effect on the overall average ratings per month between the restaurants offering and not offering takeout or delivery, thus decreasing the overall bias of the analysis.

To further validate that the quantity of review data points obtained were usable for analysis, I performed a Chi-Squared test to check if both variables, whether or not a restaurant offered takeout or delivery and whether or not they had high average ratings, were dependent. I defined high average ratings as average ratings above a 3.5 based on my personal experience as a user on Yelp and what I usually use to qualify as restaurants on the higher end of the scale. However, this has potential for some ethical concerns, as this is my own subjective opinion and others may not fully agree with my perception. For the Chi-Squared test, my null hypothesis was that the variables were independent and the null hypothesis was that the variables were not independent. First, I created a contingency table that calculated the frequency of each variable in my dataset. After performing calculations using Scipy's stats package in Python, the results showed that the test statistic of 5.481 was greater than the critical value of 3.841 and that the p-value of 0.019 was less than the level of significance of 0.05. This led to the conclusion to fail to reject the null hypothesis and conclude that there is sufficient evidence that the variables are dependent. These results validated that although Yelp limited the amount of reviews data queried per restaurant, the data points on the reviews were adequate enough to use for analysis and for assumption of dependence on the average ratings.

After additional cleansing and merging, the final dataset was produced for final analysis. For this step, I produced two pivot tables displaying the number of reviews and the average ratings for restaurants with pickup or delivery and no pickup or delivery grouped by the month. This made it simple to produce the final time series visualizations for the project. On each time series graph, there were two distinct lines comparing the two different types of restaurants (takeout or delivery and no takeout or delivery) with their respective calculated monthly values for number of reviews and average ratings. On the secondary axis, there was a bolder red line that displayed the monthly change in reported Covid-19 cases. From the produced visualizations, I noticed that the results from the time series plot comparing the number of reviews showed that for the majority of the time, restaurants offering takeout or delivery exceeded restaurants not offering takeout or delivery. Because both lines seemed to follow similar trends, I decided to produce a pie chart showing the proportions of restaurants offering takeout or delivery and those that did not in order to supplement and support the findings in this time series graph. From a human centered perspective, I think this is important because if the proportions of the types of restaurants aren't clearly displayed and made known, the times series graph would indicate that restaurants offering takeout or delivery had more reviews in total regardless of circumstance and time. However, seeing that there is a larger number of these types of restaurants helps the user understand why the graph follows this trend.

Findings

The results from the visualizations produced some interesting and surprising discoveries. One notable realization was when I first collected the data from Yelp Fusion in mid-November 2021 to produce my visualizations, the results were different from when I re-ran my Jupyter notebook later in early December 2021. In hindsight, this is as expected since the Yelp Fusion API returns a live snapshot of the status of each restaurant on Yelp at the current time queried. Especially with Yelp's default sort order for the top three reviews, the code provided makes it particularly challenging to reproduce my exact results. In this section, I will discuss and compare the resulting visualizations produced from both instances in time from mid-November 2021 and early December 2021.

Figure 1 below displays the pie chart comparing the proportion of restaurants in Wake county that offer takeout or delivery and the proportion of restaurants that do not. The figure on the left shows the results ran from the data collected in November, while the figure on the right shows the results ran from the data collected in December. From both results, you can still see that the proportion of restaurants offering takeout or delivery in Wake County are higher regardless of when the data was queried. This aligns with the increasing trend for demand in takeout and delivery options as mentioned previously in the introductory sections of this paper.

Madalyn Li
DATA 512
Fall 2021

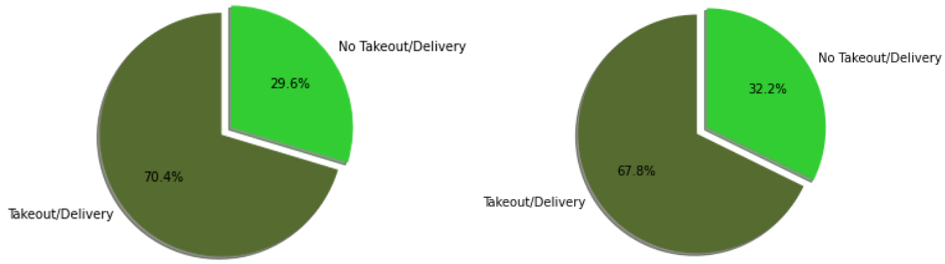

Figure 1:Proportion of restaurants with takeout or delivery and no takeout or delivery. The left chart is from November data and the right chart is from December data.

Figure 2 compares the monthly change in total reviews for restaurants offering takeout or delivery and restaurants that didn't. The resulting graph shown was produced with the data from November 2021. It is interesting to note here that for the majority of the pandemic, the total number of reviews for restaurants offering takeout or delivery almost always exceeded the total number of reviews for restaurants not offering takeout or delivery. This aligns with the previously mentioned observation of a higher proportion of restaurants offering takeout or delivery. However, one important event to make note of is that during the beginning of the pandemic, specifically between March 2020 to April 2020, there was a significant drop in the number of reviews for restaurants offering takeout or delivery. My assumption for this occurrence is that when the pandemic first hit the county, people did not feel comfortable eating outside regardless of whether or not there was a takeout or delivery option.

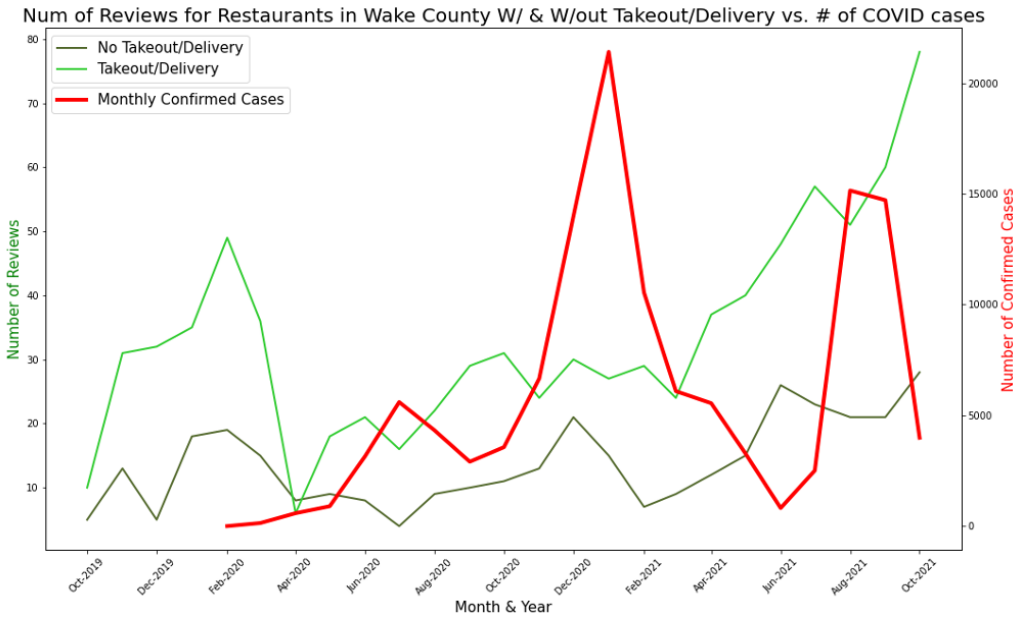

Figure 2: November 2021 Data - Monthly change in total # of reviews between restaurants with takeout/delivery and no takeout/delivery.

Figure 3 shows the same resulting graph as Figure 2, except with the data from December 2021. One additional note from this figure, is that there was missing data from July 2020 for restaurants that did not

offer takeout or delivery. This is a result of Yelp's default sort order that did not include or have any data points to match in this specific month and year. Comparing the results of this plot to the plot produced from the data in November 2021, both generally showed very similar trends for the monthly change in number of reviews. For instance, Figure 3 shows the same dip in number of reviews for restaurants offering takeout or delivery in April 2020. However, an additional significant difference in this plot is that around the peak of the second wave of reported Covid-19 cases in January 2021, there was another dip in the number of reviews for restaurants offering takeout or delivery. My assumption here is that the second wave caused a lot of fear in the county and a lot of customers also probably did not feel comfortable eating outside regardless of takeout or delivery options.

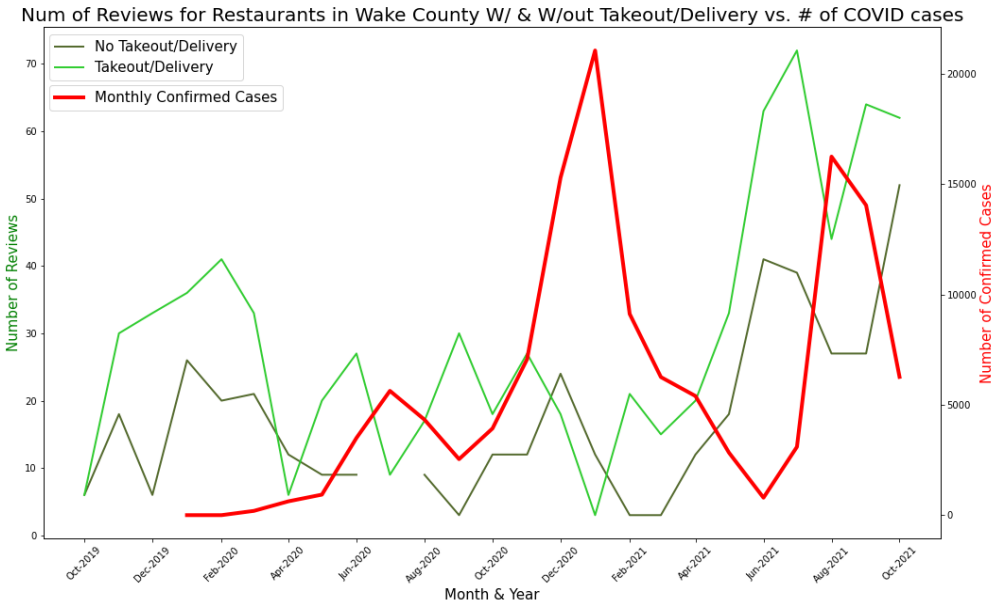

Figure 3: December 2021 Data - Monthly change in total # of reviews between restaurants with takeout/delivery and no takeout/delivery.

Figure 4 compares the monthly change in average ratings for restaurants offering takeout or delivery and restaurants that do not. The resulting graph shown was produced with the data from November 2021. The results from this time series graph indicated that there was no real trend or pattern between both types of restaurants. In fact, throughout the course of the pandemic, the restaurant type with the higher average ratings changed and shifted multiple times. I made the assumption here that regardless of whether or not a restaurant offered takeout or delivery, this made no significant impact on the monthly change in average ratings during the course of the pandemic.

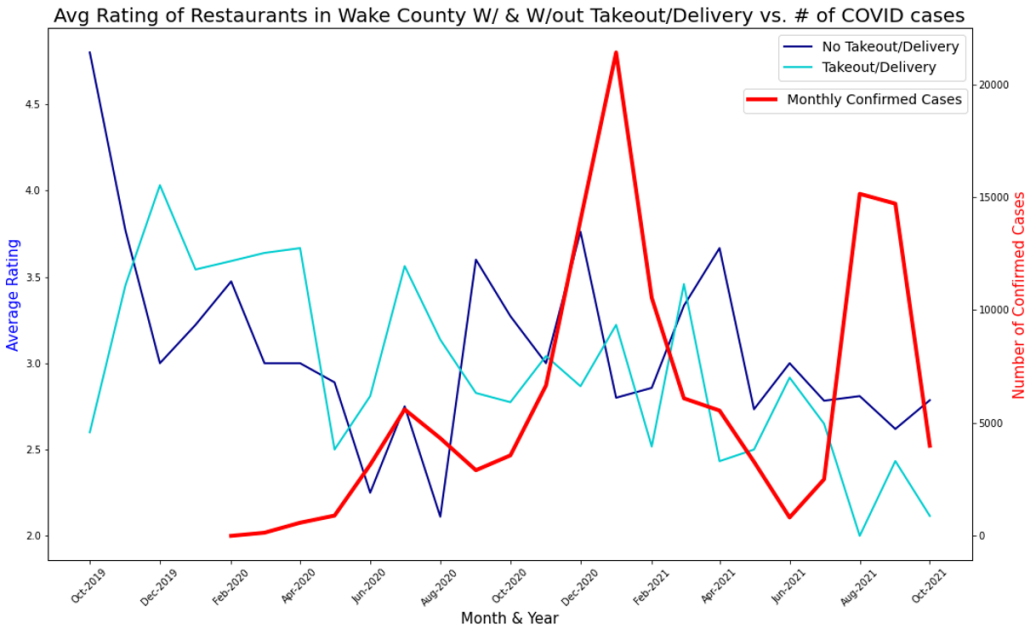

Figure 4: November 2021 Data - Monthly change in average ratings between restaurants with takeout/delivery and no takeout/delivery.

Figure 5 shows the same resulting graph as Figure 4, except with the data from December 2021. The results from this plot followed a similar trend as the previous figure and produced the same conclusion that there is no real significant pattern or trend tied to the average reviews and the takeout or delivery option offered by restaurants.

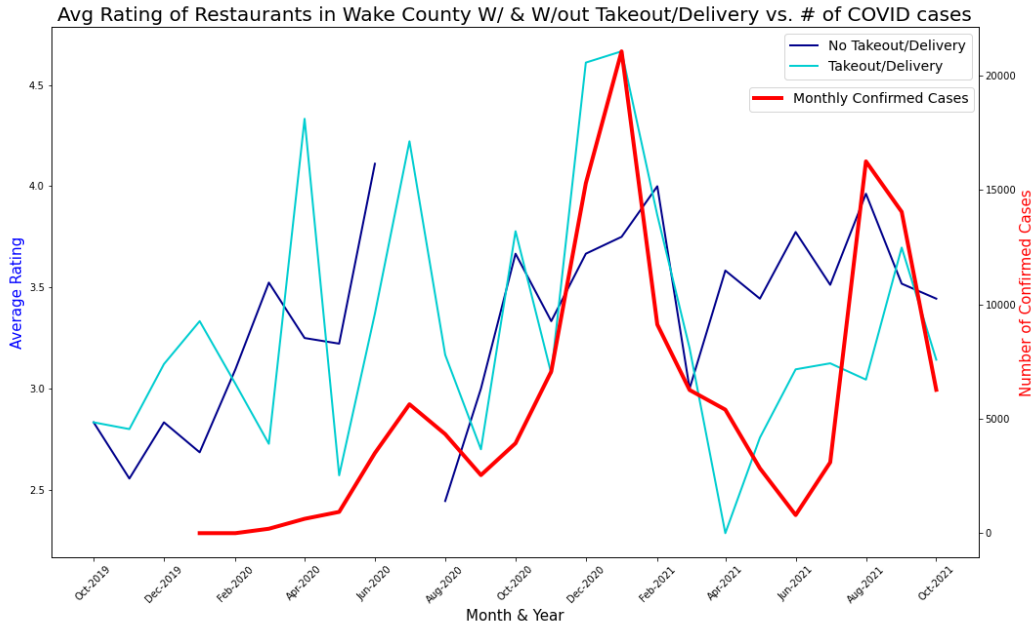

Figure 5: December 2021 Data - Monthly change in average ratings between restaurants with takeout/delivery and no takeout/delivery.

Overall, the conclusions from these plots indicated that although there are more restaurants in Wake County that offer takeout or delivery options, the impact on the monthly change in average ratings was

insignificant. This was completely different from my initial hypothesis, where I predicted to see a positive change in average ratings for restaurants offering takeout or delivery. There was, however, a general positive change in the total number of reviews for restaurants offering takeout or delivery. Nevertheless, there were also two instances of significant drops in the total number of reviews as well as missing data points on restaurants not offering takeout or delivery in July 2020 from the data queried in December which led me to believe there needs to be more data points on reviews in order to truly validate this conclusion.

Discussion/Implications

These findings are important because they provide both restaurant owners and society a glimpse into how takeout or delivery options influence society's overall perception of a restaurant. Perhaps, contrary to what surveys and articles say, it does not actually make a difference to customers whether or not a restaurant offers takeout or delivery. However, it is important to discuss that there are still some significant flaws with the data and analysis in this study. Due to Yelp's default sort order, the reviews data is constantly shifting which is problematic for reproducibility and supporting human centered data science. For future research, I would suggest collecting a more adequate and stable sample size of Yelp Reviews to validate the conclusions from this study.

The results also lead to two additional questions that future research on this subject can expand on and explore. The first question is: do Yelp reviews actually provide a good indication and measure of how society perceives the success and likeability of a restaurant? Relating this back to the concepts we studied in human centered data science, I believe this topic is important to address in order to fully understand what impacts average rating systems (which are used in many applications today) have on the perception of society. If the results show that the ratings are not a true reflection on society's beliefs and perceptions, this can help push the study to measure impact through different means. The second question is: what other factors could be incorporated or focused on to more accurately measure impact? Perhaps incorporating the distribution of age in the county or a survey measuring the general attitude of the county towards the pandemic are some suggestions to expand on the current study and measure other potential factors of influence. Additionally, it is important to note and understand that for any users reading this, the conclusion does not generalize to the entire population of Wake County or the world. This can be a dangerous misconception, but by pointing out flaws in data and analysis, transparency can provide a better understanding of current limitations for data and methodology and perhaps inspire others to improve upon it in the future.

Limitations

The most significant limitation that had the greatest impact on the results and study was the data used and derived from Yelp. As mentioned in previous sections, because the Yelp Fusion API only returned the top three reviews per restaurant, this not only limited the quantity of data points available to measure, but also severely impacted the ability to successfully reproduce results. In addition, I assumed that Yelp's default order showing the top reviews would be an accurate measure of the true ratings for the restaurant. Another limitation was the availability of data on daily transactions, orders, and revenue made by each restaurant. Although this would be a better measure for the success of restaurants during the pandemic, this data is not publicly available and is practically impossible to gather. Thus, I opted to use the data available through the Yelp Fusion API and assume that the reviews left on a restaurant's Yelp page are a good measure for likeability and popularity of a restaurant amongst the general public. Furthermore, since there is no accurate data available on Yelp to verify when a customer actually ate or ordered from a restaurant, I used the timestamps of the reviews to represent this. I assumed that there is no delay in this timestamp since I used a monthly average to calculate the number of reviews and average ratings.

Another limitation of the analysis was the difficulty to measure other factors besides a restaurant offering takeout and delivery options that influenced a customer's decision to choose a restaurant to eat at during the pandemic. There could be instances of a local hot-spot where many customers will still eat regardless of if the restaurant does not offer takeout or delivery. This external factor would be difficult to measure and calculate because it requires local knowledge and data points on heavy subjective opinions. Thus, the assumptions for this analysis was built on the fact that takeout or delivery options were a significant influencing factor on a customer's decision to eat at a specific restaurant during the pandemic. Additionally, the Yelp data available from the API does not indicate a specific time that a business began implementing takeout or delivery options. This limits results of the study because it hinders the ability to truly measure the effects of the shift to takeout or delivery since it is difficult to distinguish between the control and experimental groups that implemented these changes. Thus, from this, the assumption was made that all restaurants who currently offer takeout or delivery options, regardless of the time they implemented this procedure, was a standard measuring point as to how they were impacted during the course of the pandemic.

Conclusion

At the start of this project, the research question I set out to answer was: have restaurants in Wake County, NC that have adopted measures to conform to the shifts in demand during the pandemic, specifically

those offering takeout or delivery options, seen a change in their overall ratings and foot traffic compared to restaurants who did not choose to adapt. My hypothesis was that I would see a positive change in average ratings and number of reviews for restaurants that did implement takeout or delivery options. However, the results from my analysis indicated that regardless of a restaurant's status on offering takeout or delivery, the average ratings were not impacted.

This study builds on the understanding of human centered data science, which is an important factor to consider in any data science model or analysis. First and foremost, this study aims to answer a larger societal issue at hand by helping the reader understand how the shift in society preference due to the pandemic has impacted the restaurant industry and society's comfort level. Since data holds such significant power to provide an abundance of information and results, it is critical that as data scientists, we ensure the implications of our research and data use are for the greater good. It is also important to ensure that any bias or unfairness in our data or results are accounted for accordingly and documented properly. Finally, this study provides a great example of presenting full transparency and documentation of research and assumptions, which is equally as important to consider and include as a data scientist. Without these considerations, users have the potential to be misinformed on specific information and details, which can lead to further spread of misinformation to the general public and dangerous repercussions to society.

References

1. https://www.forbes.com/sites/aliciakelso/2020/07/07/the-covid-19-crisis-will-likely-lead-to-a-massive-shakeout-in-the-restaurant-industry/?sh=5f584efb42a0
2. https://restaurant.org/nra/media/downloads/pdfs/business/covid-19-restaurant-impact-survey-september-2021.pdf
3. https://restaurants.yelp.com/articles/dine-in-vs-takeout-trends/
4. https://www.nytimes.com/2021/07/02/technology/restaurant-delivery-takeout-orders-covid-coronavirus.html
5. https://github.com/nytimes/covid-19-data/tree/master/mask-use

Data Sources

1. https://data-wake.opendata.arcgis.com/datasets/Wake::restaurants-in-wake-county-yelp/about
2. https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
3. https://www.yelp.com/developers/documentation/v3