R·I·T

SAUNDERS
College of BUSINESS

An Analysis of Wegmans' Customer's Grocery Shopping Habits During the COVID-19 Pandemic

Authors:
Dana DiVincenzo
Devren Purdie
Pyone Myat Maw
Surbhi Sharma
Grace Matabishi

Supervisor:
Dr. Victor J Perotti

A Report

Submitted to the Department of MIS, Marketing & Digital Business

In Fulfillment of the Requirements

For the Course of BANA.785

Project Overview

Wegmans Food Markets is a supermarket chain which was founded and is headquartered in Rochester NY. Besides selling grocery items, Wegmans has expanded into selling prepared, hot meals such as subs, pizza, buffet items, and more. Wegmans also operates in-store restaurants such as their “Burger Bar” in their Perinton & Pittsford locations.

When the COVID-19 pandemic hit the United States in early 2020, consumers found that they needed to change their shopping habits. Social distancing guidelines, quarantine restrictions, and restaurant shutdowns forced many people to take fewer trips to the grocery store, and to stock up on essential items.

Partnering with Wegmans, and using the consumer and transactional data provided for Monroe county locations, our goal was to investigate how Wegmans’ consumer habits changed at the onset of the COVID-19 crisis and beyond. Furthermore, we specifically decided to focus on analyzing how restaurant shutdowns in the area may have affected sales in different departments.

Data

Wegmans data

We were provided with 4 data files from Wegmans containing data about their customers, transactions and products on the shelves at Wegmans. The data file were as follows:

Sales Data: Containing data for each transaction from January 2019 to January 2021. The variables in this file were namely Date of transaction, Transaction Id, Item number, Customer Id, weight, unit, and Sales

Item Data: This file contained information about each product/item sold at Wegmans and which class, category and department the products belonged to. The major variable names were- Item number, Item description, Department name, Category name, Class name and Product Hierarchy.

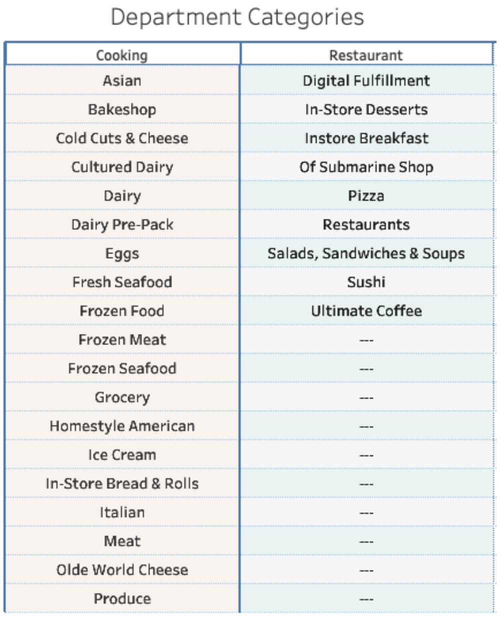
Customer Data: This data file contained the general information about each customer at Wegmans such as age, income, and number of children in the household. The variable names in the file are as follows: Customer Id, Household Id, Household Age, Income Category and number of children.

Instacart Data: Wegmans uses the Instacart app to deliver groceries to their customers. This data file contains each transaction at Wegmans and describes whether it was made through Instacart and at which Wegmans location in Monroe county the order was placed. The variables are - Transaction Id, Location number and Instacart_Ind(boolean variable).

Additional data

As our goal was to observe the Wegmans’ customer shopping habits during Covid-19 pandemic, we collected additional information such as Covid-19 positive daily cases for Monroe County starting from March 2020 to consider it as an independent variable when we run the predictive models to oversee the future sales.

Hypothesis and Assumptions



Department Categories

Cooking	Restaurant
Asian	Digital Fulfillment
Bakeshop	In-Store Desserts
Cold Cuts & Cheese	Instore Breakfast
Cultured Dairy	Of Submarine Shop
Dairy	Pizza
Dairy Pre-Pack	Restaurants
Eggs	Salads, Sandwiches & Soups
Fresh Seafood	Sushi
Frozen Food	Ultimate Coffee
Frozen Meat	---
Frozen Seafood	---
Grocery	---
Homestyle American	---
Ice Cream	---
In-Store Bread & Rolls	---
Italian	---
Meat	---
Olde World Cheese	---
Produce	---

Due to the pandemic, the NYS government placed restrictions on restaurants in the state. These restrictions changed over time, due to many variables but the greatest one being the infection rate of an area. On 3/16/2020, restaurants in Monroe County were forced to shut down all services besides takeout. On 6/4/2020, restaurants could offer outdoor seating, and on 6/12/2020, restaurants could offer indoor seating, with various restrictions such as capacity limitations and curfews. This limitation continued for the rest of 2020.

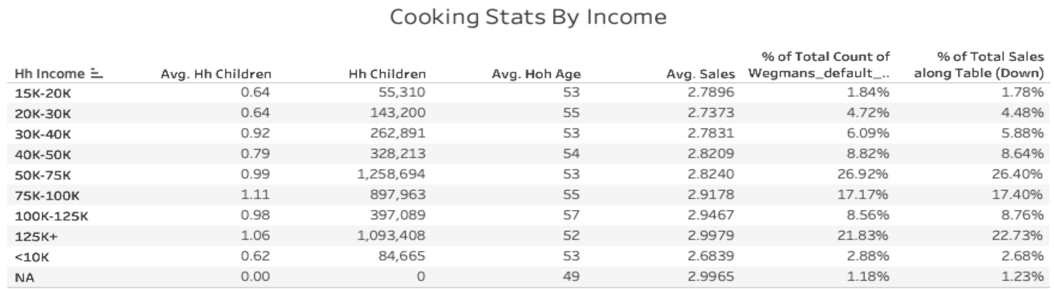
We believe that due to restaurant shutdowns, consumers who often ate out at restaurants would be forced to cook more at home. Therefore, we hypothesize that Wegmans should have seen an increase in sales of foods that fall into “cooking” departments. In tandem, sales of items within “restaurant” departments should have decreased, as Wegmans was prohibited from offering many types of prepared foods at that time. As restaurant restrictions loosened, consumers will likely return to their old habits and sales in the “cooking” and “restaurant” departments would get closer to how they were pre-pandemic. The determination of department categories is shown in the above table.

Hypothesis Objectives

Our objective is to investigate how sales in “cooking” and “restaurant” departments changed during the onset of COVID-19, and how they behaved after the initial wave of cases and restrictions. If changes are determined to be significant and long-lasting, Wegmans can use this data to better determine how to react in any potential future waves of infections in the area.

Exploratory data analysis (EDA)

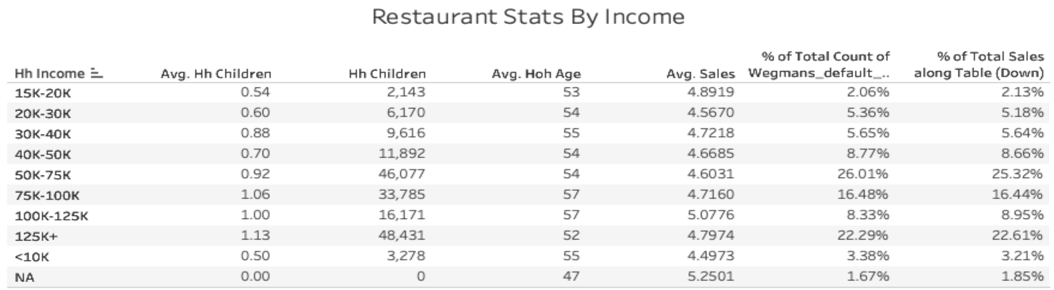
Statistical Data insights



Cooking Stats By Income

Hh Income	Avg. Hh Children	Hh Children	Avg. Hoh Age	Avg. Sales	% of Total Count of Wegmans_default_..	% of Total Sales along Table (Down)
15K-20K	0.64	55,310	53	2.7896	1.84%	1.78%
20K-30K	0.64	143,200	55	2.7373	4.72%	4.48%
30K-40K	0.92	262,891	53	2.7831	6.09%	5.88%
40K-50K	0.79	328,213	54	2.8209	8.82%	8.64%
50K-75K	0.99	1,258,694	53	2.8240	26.92%	26.40%
75K-100K	1.11	897,963	55	2.9178	17.17%	17.40%
100K-125K	0.98	397,089	57	2.9467	8.56%	8.76%
125K+	1.06	1,093,408	52	2.9979	21.83%	22.73%
<10K	0.62	84,665	53	2.6839	2.88%	2.68%
NA	0.00	0	49	2.9965	1.18%	1.23%

There is not much difference in who buys foods from cooking departments vs those who buys from the restuarant departments. Children, average age and % of sales are all the same. Though, there is an increase in average sales per customer when they purchase from restaurant departments. This makes a lot of since because all restaurant items are higher in price to account for labor and packaging costs.



Restaurant Stats By Income

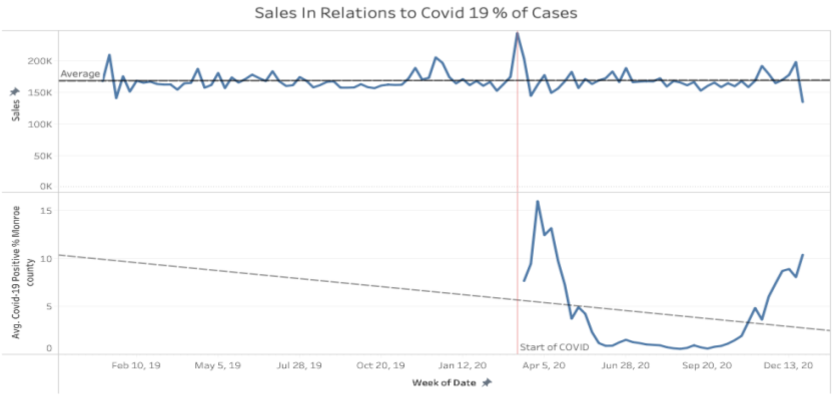
Hh Income	Avg. Hh Children	Hh Children	Avg. Hoh Age	Avg. Sales	% of Total Count of Wegmans_default_..	% of Total Sales along Table (Down)
15K-20K	0.54	2,143	53	4.8919	2.06%	2.13%
20K-30K	0.60	6,170	54	4.5670	5.36%	5.18%
30K-40K	0.88	9,616	55	4.7218	5.65%	5.64%
40K-50K	0.70	11,892	54	4.6685	8.77%	8.66%
50K-75K	0.92	46,077	54	4.6031	26.01%	25.32%
75K-100K	1.06	33,785	57	4.7160	16.48%	16.44%
100K-125K	1.00	16,171	57	5.0776	8.33%	8.95%
125K+	1.13	48,431	52	4.7974	22.29%	22.61%
<10K	0.50	3,278	55	4.4973	3.38%	3.21%
NA	0.00	0	47	5.2501	1.67%	1.85%

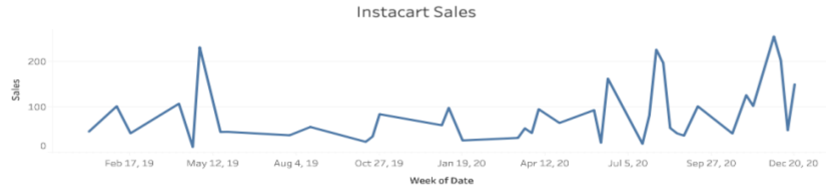
As we began to investigate the differences between the cooking categories and restaurant categories, we wanted to know just how different they are. We decided to break up the data into two separate sets; data that was only from the cooking categories, and data from the Restaurant categories. We then looked at several factors within each subset; the average number of children in the household, the total number of children, Head of household average age, the average sales per category, row count and how much that category of income contributed towards total sales.

We found that there are far more transactions for cooking products opposed to restaurant products overall. An interesting insight was to realize that there wasn't much of a difference between the income categories. A customer in the $15k-20k income range creates almost the same amount of sales dollar per transaction as a customer in the $125k+. The average age of customers consistently falls between 52-57 and the number of children per household is pretty much 1 on average.

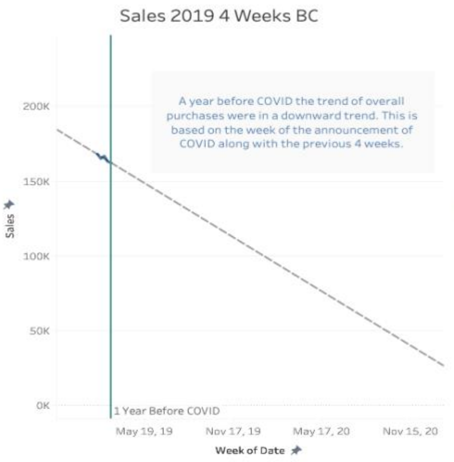
Visualization insights

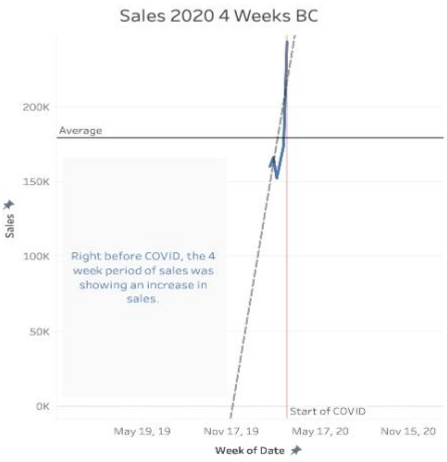
1) General findings





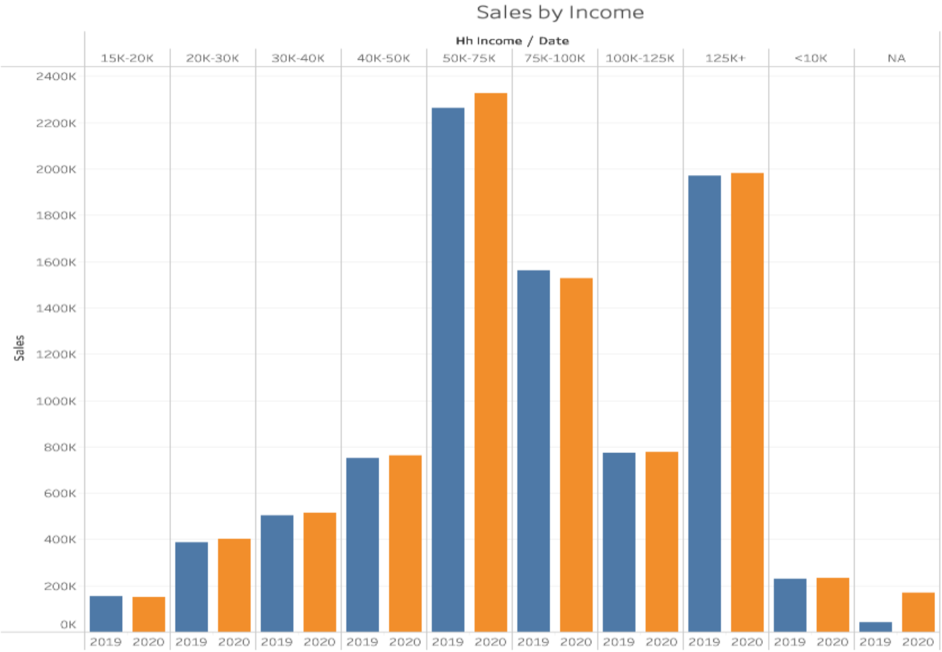
According to the data provided to us, we see that the sales had a significant increase in March 2020 when the pandemic first started having a spike in cases as the restrictions followed, however, this increase dropped back to normal sales pattern by the end of April 2020. There was no other abnormality in sales at all. It is important to note that Instacart has been gaining steady steam for months before COVID but was granted a generous growth curve after COVID took the country by storm.

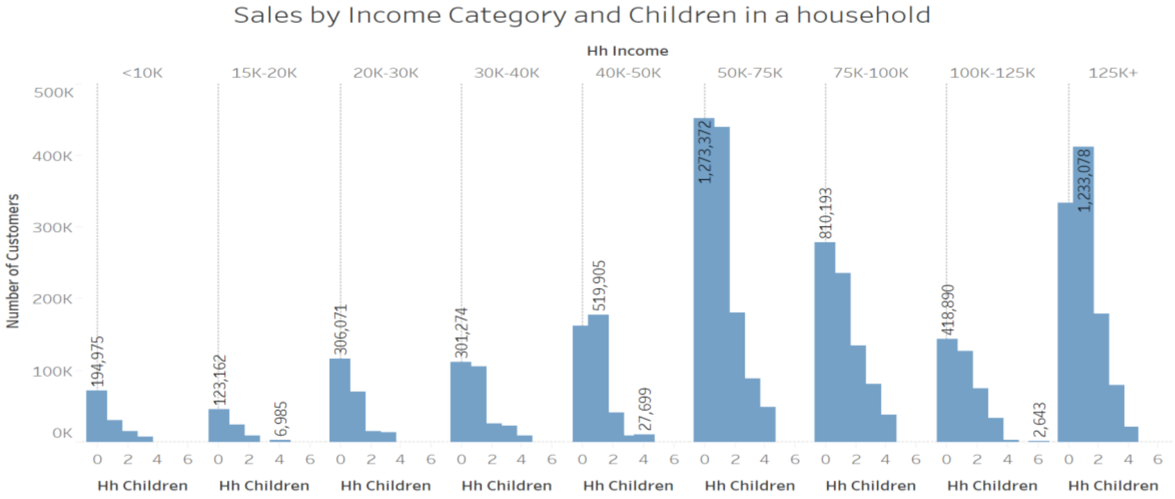




Last year's sales are the best indicator to this year's sales. Because COVID was the only thing that altered the sales pattern, we wanted to investigate just how much COVID affected sales for the weeks leading up to COVID. We took a peak at the 4 week trend of sales for the week of COVID from 2019. A downward trend is the pattern for sales as we entered March of 2019, compared to a steep slope upwards in 2020 as customers began to feel the pressure of COVID 19 picking up.

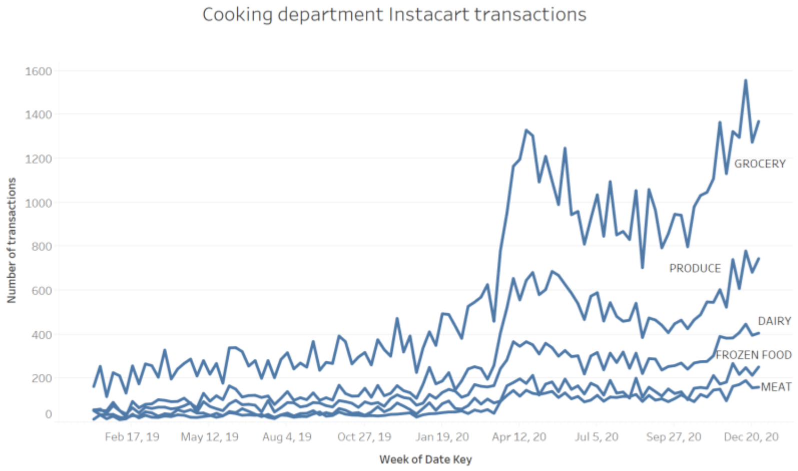
The sales in total remained quite like previous year for different income categories, leaning us towards the possibility that covid did not impact the overall sales for Wegmans. This is proving that the increase in sales seen in March was the only abnormality in sales. When considering two years of data, there wasn't really any change in sales from 2019 to 2020.





The graph above tells us how many customers have the specified number of children across each income category along with the sum of sales labels for the highest frequency of customers. Most of the customers belong to the 50k-75k and 125k+ income category. One of the reasons why 50k - 75k has the greatest number of customers could be because the average household income for Monroe county was $60,075 (According to U.S. census, 2019) which falls in this category. Within these categories, the frequency of customers with zero or less than 2 children at Wegmans is much more compared to households having 2 or more than 2 children. One would assume that people with a family (in this case- 2 or more than 2 children) would spend more on groceries, however, this data doesn't support that thought.

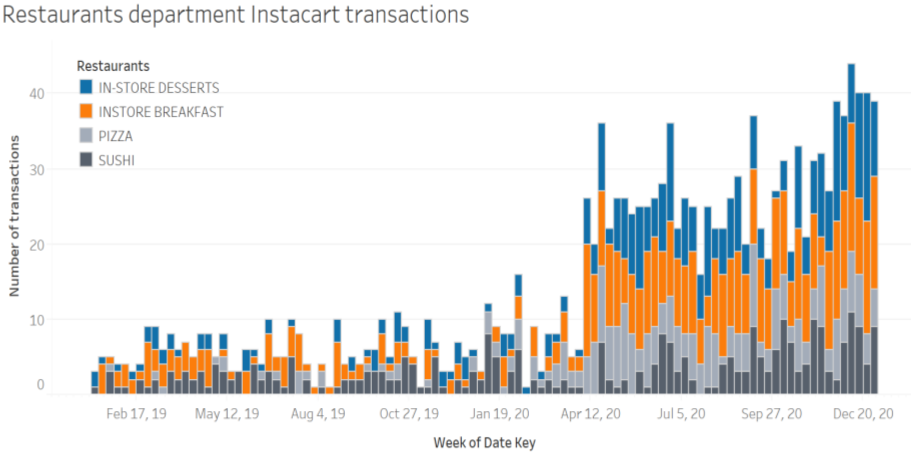
2) Hypothesis findings



After deciding to go for two different department categories - Cooking and Restaurant, we wanted to see if the Instacart variable had any significance in department sales. Our assumption was that if people were ordering from the app (Instacart), that they were ordering groceries or in our case - the cooking department products. We also wanted to see if online ordering increased due to the pandemic, and if it did, which departments saw the highest growth in sales.

The above graph shows the total Instacart transactions for the top 5 cooking departments according to sales. There is a spike in online orders for the cooking departments. The sales spiked in March and April 2020 but did not go down to their original trend of sales. There was a growth in online orders during the pandemic. This also tells us that after the start of the pandemic, people trust Instacart a lot more and are using it to avoid going to the store for their shopping.

The graph shows the number of Instacart transactions for the top 4 restaurant departments. We saw a spike in online transactions for these categories in April and even after many COVID restrictions were lifted in June 2020, the Instacart transactions stayed up compared to transactions before COVID. In addition, we notice that the Pizza department had a significant increase in orders during COVID compared to other categories.



Predictive Modeling & Extensive Analysis

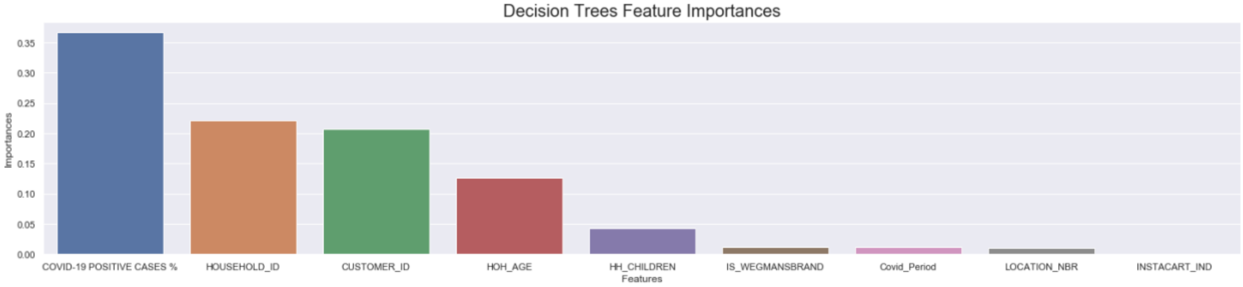
Decision Trees Model

Using the decision trees model we were able to generate the features' importance and relationships between dependent and independent variables.

1) Data Cleaning and Preprocessing:

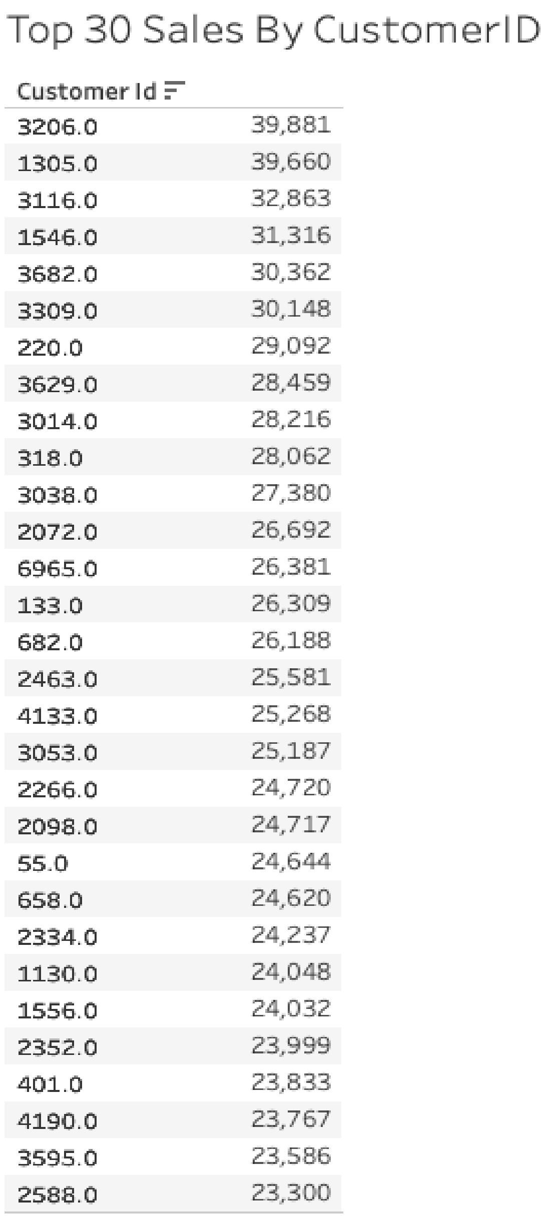
Python programming language was used to do data cleaning and preprocessing to run the decision tree model on the selected features from the dataset. Firstly, the classification groups as cooking, restaurants, and others based on department categories according to our hypothesis and its objectives for this analysis. Then the collected Covid 19 % positive cases for Monroe County was merged with the Wegmans' dataset that includes the information of daily sales for each department, customer ID and insta cart. The new variable was also created for defining the before and during Covid-19 periods as 0 and 1 (0- before and 1- during Covid). The missing values of datasets were dealt differently for categorical variables, numerical variables, and other variables by replacing them with "NA", "average", and "0". We then filtered the sales that belong only to the cooking categories to run the decision tree model and the nine features shown in the below bar chart in the feature importance session were selected to see their correlations with the fluctuation of sales for two years of 2019 and 2020.

2) Feature Importance:



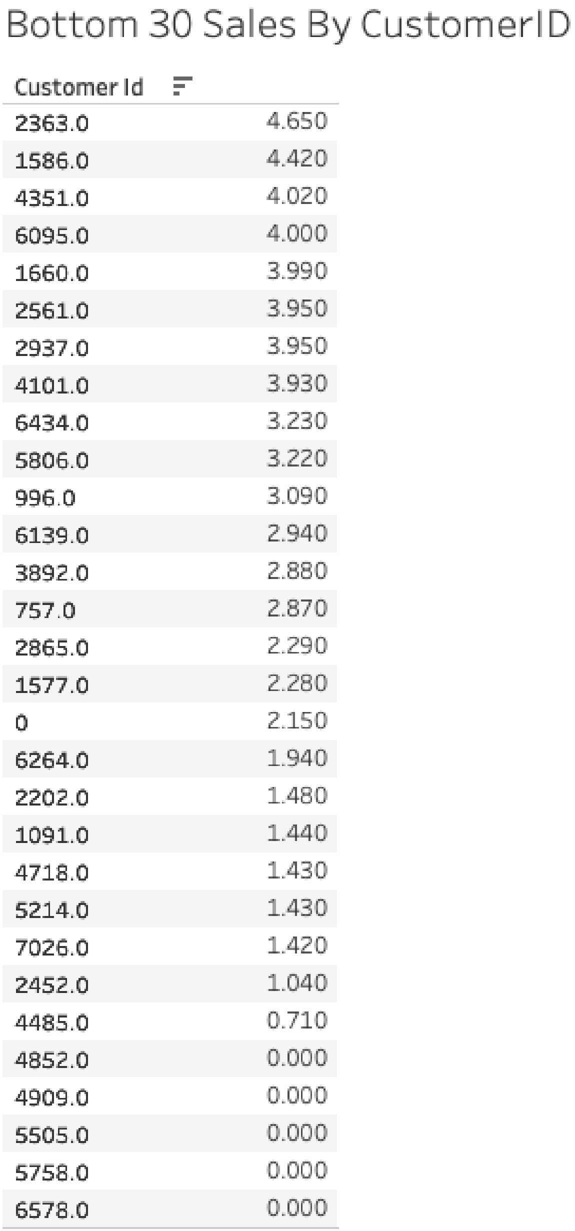
This bar chart shows the importance of features that have a high influence on the sales. It shows that the Covid-19 positive cases feature scores as a top important feature compared to the rest. This result is also inline with the trends seen in EDA session as the sales in 2020 had a few spike periods while Covid-19 cases were obviously going up.

According to the datasets given by Wegmans, it is also assumed that customer ID and Household ID would be the same and given the result in the bar chart reveals that the specific customer could be next highest influencing on the sales. To study the relationship between customer and sales, the following statistics have been conducted to see the total sales made by top and bottom 30 customers according to their aggregate spending from 2019 & 2020. It can be said that the average repeat/loyal customers would spend a lot more than the new/normal customers.



Top 30 Sales By CustomerID

Customer Id	
3206.0	39,881
1305.0	39,660
3116.0	32,863
1546.0	31,316
3682.0	30,362
3309.0	30,148
220.0	29,092
3629.0	28,459
3014.0	28,216
318.0	28,062
3038.0	27,380
2072.0	26,692
6965.0	26,381
133.0	26,309
682.0	26,188
2463.0	25,581
4133.0	25,268
3053.0	25,187
2266.0	24,720
2098.0	24,717
55.0	24,644
658.0	24,620
2334.0	24,237
1130.0	24,048
1556.0	24,032
2352.0	23,999
401.0	23,833
4190.0	23,767
3595.0	23,586
2588.0	23,300



Bottom 30 Sales By CustomerID

Customer Id	
2363.0	4.650
1586.0	4.420
4351.0	4.020
6095.0	4.000
1660.0	3.990
2561.0	3.950
2937.0	3.950
4101.0	3.930
6434.0	3.230
5806.0	3.220
996.0	3.090
6139.0	2.940
3892.0	2.880
757.0	2.870
2865.0	2.290
1577.0	2.280
0	2.150
6264.0	1.940
2202.0	1.480
1091.0	1.440
4718.0	1.430
5214.0	1.430
7026.0	1.420
2452.0	1.040
4485.0	0.710
4852.0	0.000
4909.0	0.000
5505.0	0.000
5758.0	0.000
6578.0	0.000

Market Basket Analysis

While market basket analysis is an unsupervised technique that offers no predictions, it does allow us to discover relationships between items within sets. Here, we applied market basket analysis techniques to our department level data, rather than items. Our goals with our analysis here were to:

1. See if customers who bought items from cooking departments were more likely to buy items from other cooking departments compared to restaurants or other departments.
2. See how the probability of (1) changed when the pandemic began and afterward.

1) Data Preparation:

The data from the sales and item datasets were merged to make a dataframe with only the transaction key, the department name, and the date of the transaction. The data was cleaned and the "bottle container returns" department was removed, as it is not relevant to this analysis. The data was then split into 3 periods:

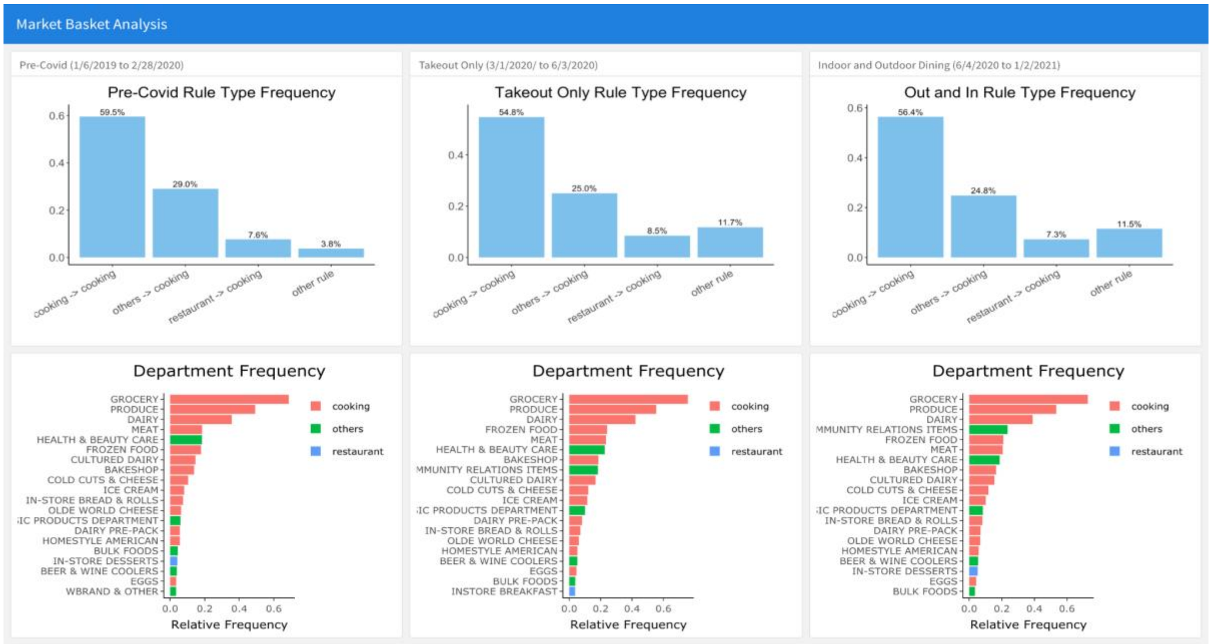
1. Pre-pandemic: 1/6/2019 to 2/28/2020.
2. Early pandemic / restaurants were restricted to takeout only: 3/1/2020 to 6/3/2020.
3. Mid pandemic / restaurants were open again with some restrictions: 6/4/2020 to 1/2/2021.

2) Data Mining:

The same analysis was run on all three data frames listed above.
First, the top 15 departments were found for each dataset. Next, using the apriori algorithm, market basket rules were created for departments within the data frame. The max rule length was set at 2, and the confidence of the rule was set at 25%. Using the breakdown of department categories defined at the beginning of this report, the created rules were classified into rule types. The possible rule types were:

1. Cooking department item implies purchase of another cooking department item.
2. Restaurant department item implies purchase of a cooking department item.
3. Other department items imply purchase of a cooking department item.
4. Other types of rules (such as restaurant implies restaurant).

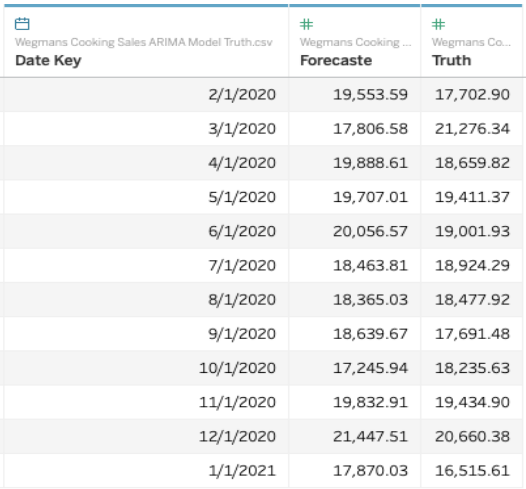
The top 15 items and breakdown of rule types were all displayed graphically using flexDashboard in rStudio. These visualizations are shown below.



As shown in the figure above, the first bar in the top rows of graphs shows us the percentage of rules which were ‘cooking department item implying the purchase of another cooking department item’. Based on our hypothesis, we expected to see this percentage increase significantly from the first graph to the second graph, or when COVID-19 hit Monroe County. However, we can see that this percentage decreased slightly at this time and increased slightly in the later period. This tells us that there was not a significant change in the types of departments consumers purchased from when the pandemic hit. The figures in the bottom row show us the departments which were purchased from most frequently and are broken down further into their department category. All these images tell us a similar story - that cooking departments are purchased from the most, with other departments mixed in towards the top as well. Restaurant departments hovered near the bottom for all three time periods.

Autoregressive integrated moving average (ARIMA)

As time series analysis is reliable for analyzing time series data in order to observe the meaningful statistics and characteristics of the data, the ARIMA time series model was conducted to predict the following future values based on previously observed sales. ARIMA model could help us to generate forecasting sales for 10 years (2019 to 2029).



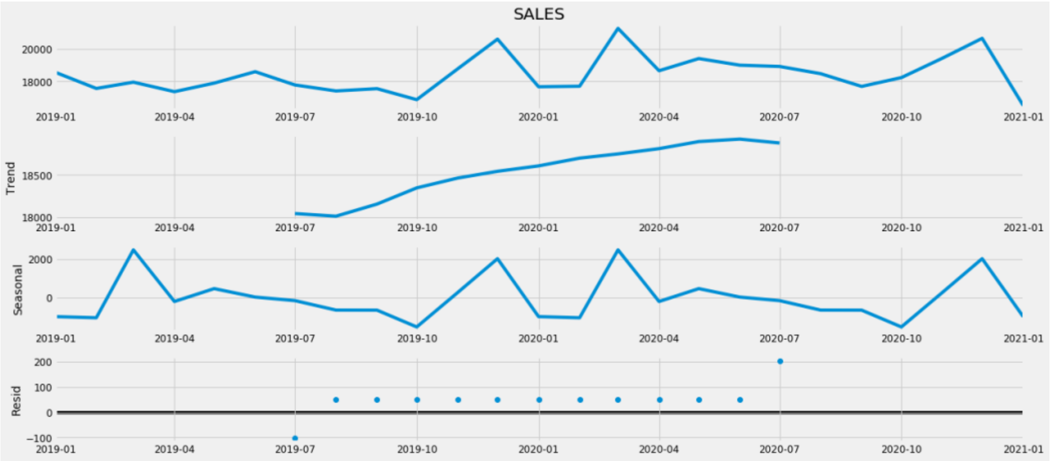
Wegmans Cooking Sales ARIMA Model Truth.csv Date Key	Wegmans Cooking ... Forecaste	Wegmans Co... Truth
2/1/2020	19,553.59	17,702.90
3/1/2020	17,806.58	21,276.34
4/1/2020	19,888.61	18,659.82
5/1/2020	19,707.01	19,411.37
6/1/2020	20,056.57	19,001.93
7/1/2020	18,463.81	18,924.29
8/1/2020	18,365.03	18,477.92
9/1/2020	18,639.67	17,691.48
10/1/2020	17,245.94	18,235.63
11/1/2020	19,832.91	19,434.90
12/1/2020	21,447.51	20,660.38
1/1/2021	17,870.03	16,515.61

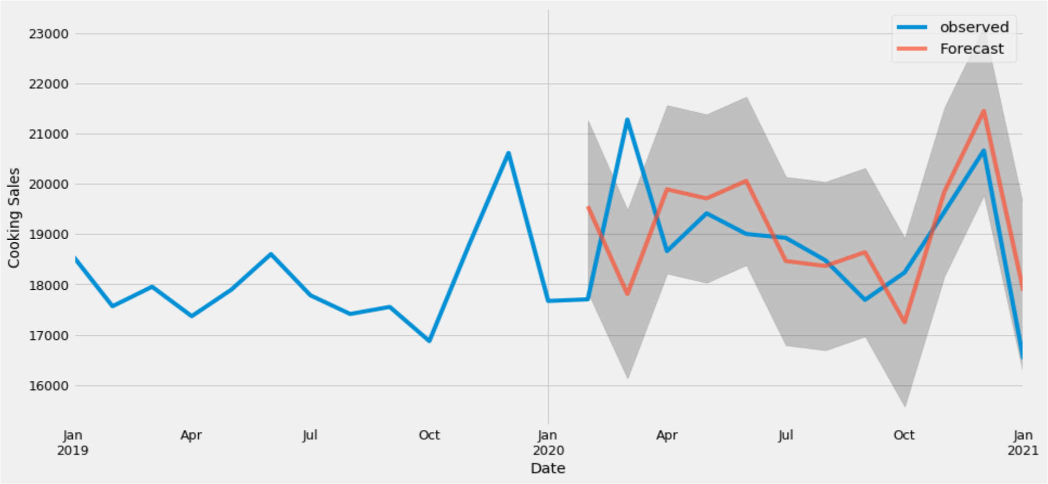
1) Data Cleaning and Preprocessing:

This step included removing columns that are not needed to run the model and missing values were dealt as we did in the decision trees model. The sales were aggregated by date. As the given Wegmans datetime data could be tricky to work with, we used the average daily sales value for each month instead and using the start of each month as the timestamp.

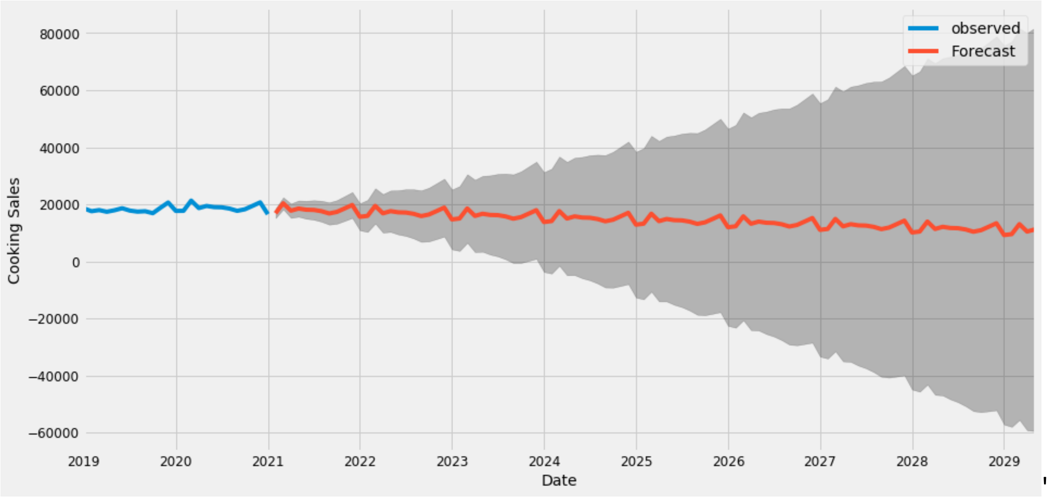
2) Predictive data visualizations

Then, the time series data were visualized by using a method called time-series decomposition that allows us to divide our time series into three distinct components as tren, seasonality, and residuals which can be seen in the following images.





This graph maps out the ARIMA model prediction. The model did show very similar results to the predicted sales. The largest difference was the spike in sales from COVID. This reinforces the fact that last year sales will always be the best indicator of this year's sales.

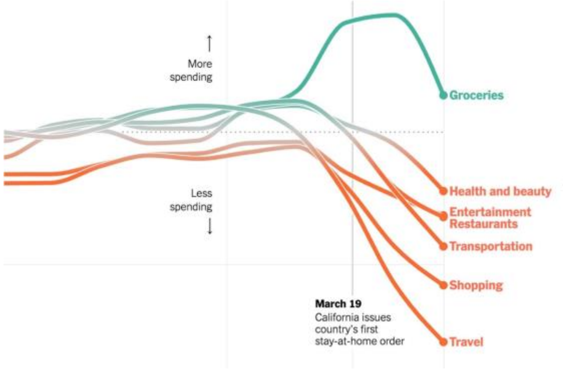


Future Scope

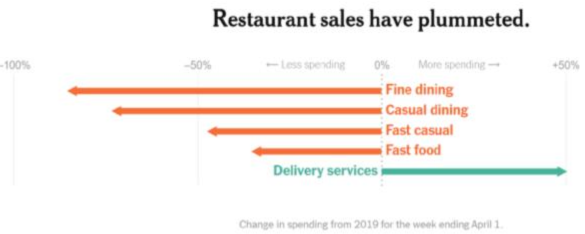
After the above predictive models and analysis have told us about the future sales trends and the relationships between the variables, we realized that there are some techniques that we can apply to continue studying the Wegmans' consumer's behaviours that could impact the future sales.

Customer lifetime value analysis could help us to determine the values and churn of customers while cluster analysis can give us more insightful information about each determined group of customers so that we can achieve better effective customer marking. By making and setting rules and thresholds in decision trees the models would reveal the if there are relationships between the variables. For instance, if a customer who has 3 children spends more than $100 on a trip, would that lead to an increased likelihood of another variable. Another way to dig deeper into the customer shopping habits to find out how customers felt about Wegmans during the pandemic. We can do this by performing a sentiment analysis based on the posts, tweets or stories being posted on social media. We can analyze if customers are more frustrated with Wegmans or if they are happy with their in-store (or online) experiences since the start of COVID.

Conclusions



March 19
California issues country's first stay-at-home order

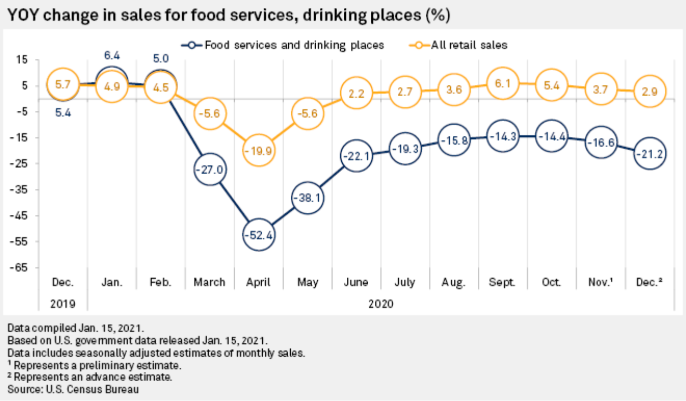

Restaurant sales have plummeted.

Change in spending from 2019 for the week ending April 1.

The above shown industry trends gives us an overview of how the food industry had a different sales pattern during Covid than other industries. Compared to other industries, groceries had more spending from customers throughout the weeks surrounding COVID 19 while the other industries took a hit in sales due to Covid. Our analysis of Wegmans data proves the hypothesis that grocery sales (in our case, cooking department sales) experienced a major increase in sales in the first week of March and then the sales were back to their normal pattern by mid 2020 whilst maintaining their previous year sales pattern.

We also notice that restaurant sales have plummeted all over the country, however, delivery services have seen a major increase in sales. Similarly, Wegmans went through a decrease in restaurant sales in the first few months of Covid but resumed to their normal sales trends of 2019 after June 2020. Wegmans evidently have been doing much better than overall industry patterns.

In line with delivery services having more sales, Wegmans also experienced an increase in instacart sales in 2020 as compared to 2019 in both cooking and restaurant departments.


YOY change in sales for food services, drinking places (%)

Data compiled Jan. 15, 2021.
Based on U.S. government data released Jan. 15, 2021.
Data includes seasonally adjusted estimates of monthly sales.
[1] Represents a preliminary estimate.
[2] Represents an advance estimate.
Source: U.S. Census Bureau

When restaurants closed due to mandates, the restaurant industry saw an average of a 52.4% decline in sales. The trend for 2020 stayed around a 20% loss of sales, but for Wegmans, there was only a dip in sales from March through June. For more than half of the year most restaurants suffered with significant losses while Wegmans continued to reap profits at normal ratios. Compared to their competitors, Wegmans did exceptionally well. These facts emphasize the fact that Wegmans customers did not change their buying habits. The loyalty of Wegmans customers gave Wegmans no reason to worry about COVID 19. We predict that if another pandemic or national emergency were to take place, that Wegmans would stay right where they are.

Sources

“Documenting New York's Path to Recovery from the Coronavirus (COVID-19) PANDEMIC,

2020-2021.”

Ballotpedia, 2021,

ballotpedia.org/Documenting_New_York%27s_path_to_recovery_from_the_coronavirus_(COVID-19)_pandemic,_2020-2021.

Leatherby, Lauren, and David Gelles. “How the VIRUS Transformed the Way Americans Spend

Their Money.” *The New York Times*, The New York Times, 11 Apr. 2020, www.nytimes.com/interactive/2020/04/11/business/economy/coronavirus-us-economy-spending.html.

Michael O'Connor, Chris Hudgins. “Dining out: DECEMBER Restaurant Sales Decline, Capping off a Year of Crisis.” *Accelerating Progress*, 15 Jan. 2021, www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/dining-out-december-restaurant-sales-decline-capping-off-a-year-of-crisis-62120179.

Staff, WROC. “86 New COVID-19 Cases in Monroe County, Largest Single Day Increase since May 19.” *RochesterFirst*, RochesterFirst, 22 Oct. 2020, www.rochesterfirst.com/coronavirus/86-new-covid-19-cases-in-monroe-county-largest-single-day-increase-since-may-19/.