MarginEdge Pulse:

Predicting Restaurant Sales

Bo Davis: bo@marginedge.com,
Julie Davis: juliedavis80@gmail.com,
Yabo Gao: yabo.nancy.gao@gmail.com,
Chelsea Weiss: chelsea.l.weiss@gmail.com

8315 LEE HWY, SUITE 320, MERRIFIELD, VA 22031
888.488.9612

Contents

Executive Summary

As the old saying goes, “hindsight is 20/20.” While humans cannot always accurately predict the future, we can look back and realize how much better we could have done if we knew then what we know now. This phrase applies well to the restaurant industry. If restaurants were to know in advance what their customers want, they would be able to do a much better job of maximizing their revenue and reducing cost. MarginEdge Pulse is MarginEdge’s proposed product offering that will be its restaurants’ crystal ball into the future. The company teamed up with a group of researchers from Northwestern University to build a Data Science framework to take the nightly data feeds and make reliable sales predictions. The predictions will be per restaurant, per day, with a seven-day rolling look ahead (adjusted nightly).

About MarginEdge

MarginEdge is a leading software company providing a platform that integrates with point-of-sale (POS) systems for restaurants to manage more efficiently their "back of house" operations like purchasing, inventory management, and labor. MarginEdge currently works with 2,000+ restaurants around the country and is growing rapidly – doubling clientele numbers each year.

Problem Statement

Restaurants are famously challenging to manage profitably. At the heart of the issue are high variable costs such as food and labor that require spending for purchasing and scheduling before the sales are known. This issue requires a sales estimate and pre-spending against that target. For example, if you expect a $5,000 day in sales, you may schedule $1,500 in labor and buy $1,500 worth of food. However, if your sales come in at $4,000, you are throwing away valuable food and wasting labor – in what is already a very thin margin business! On the other hand, if you under-purchase food or labor because your sales are higher than expected, you have significant food quality and service issues.

Pair these inherent challenges with the dining crisis caused by the novel Covid-19 global pandemic, and restaurants have struggled to maintain operation with desperately slim margins. Getting the cost ratio right requires a lot of guesswork on the restaurant operator's part, but it is a problem MarginEdge can solve for them!

Strategic Value

In MarginEdge's software, a core feature is to allow restaurant managers to track variable costs against targets. Figure 1 illustrates a sample dashboard that helps a restaurant manager see the targeted versus actual trend and the estimated next day. Although a next day prediction is useful, it would be much more beneficial to our clients to have one-week sales predictions by day. This enhancement will enable managers to determine the levels at which they should purchase inventory and schedule staff more efficiently. When the restaurant operator is placing their purchases through MarginEdge's proprietary software, we will use this *predicted sales* value to show purchases against this target. The predictive power of future forecasting allows MarginEdge's clients to save money and increase their profit margin.

Figure 1. Example of MarginEdge's "25% to Target Spend" dashboard

Literature Review

Businesses have often struggled with the crystal ball problem. If only they knew what the future held for their sales, they could accurately plan for inventory and staff to maximize their profit margins. This issue is detailed in research as companies struggle with predicting and planning for supply chain and sales fluctuations (Chen et al. 2018, 56). Restaurants suffer acutely from this dilemma because of their inherently small profit margins. Many data scientists have detailed this issue, saying, "forecasting future demand is vital to planning and operations in the restaurant industry at both the micro and macro levels" (Reynolds, Rahman, and Balinbin 2013, 317). Reynolds, Rahman, and Balinbin further noted, "Extending this to restaurant suppliers and distributors, simple common sense tells us that being able to forecast expected growth in aggregate restaurant sales well in advance should help operators in planning, modernizing, or constructing facilities, hiring employees, and developing other related assets to ensure commensurate growth in services, revenue, and competitive advantage. Extended one step further, accurately forecasted aggregate sales will assist both the individual and large institutional investors, as well as lenders, in making the 'go, no-go' decision" (2013, 317).

In response to this need, Reynolds, Rahman, and Balinbin conducted a study to "explore likely predictor variables that should be integrated into industry-specific econometric models and to test the viability of these variables across subsegments" (2013, 317). The researchers considered "general predictor variables such as percentile change in the consumer price index (CPI), percentile change in food expenditures outside the home, percentile change in population, and percentile change in the unemployment rate to see whether an econometric model can indicate their effects on aggregate annual restaurant sales lagged for one year." (2013, 318). The data spans from 1970 to 2011, and Reynolds, Rahman, and Balinbin "applied

the model independently to full-service restaurant sales, quick-service restaurants sales, on-site foodservice restaurant sales, non-commercial restaurants, and what we call drinking places sales" (2013, 318).

Reynolds, Rahman, and Balinbin developed a single-equation econometric model that connects a dependent variable to several distinct input variables (2013, 318). The equation is as follows:

$$Y_t = \beta_o + \sum_{1}^{J} \beta_i X_{it} + e_t,$$

The authors of this study developed several models for this problem in search of the most optimal solution. The results are shown in Table 1 (2013, 321).

	Full service		Quick service		On site		Drinking		Non-commercial	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Model 1		.400**		.006		.368**		.370**		.404**
Δ CPI	−.632**		−.080		−.607**		−.608**		−.636**	
Model 2		.105**		.059		.101*		.097*		.092*
Δ CPI	−.578**		−.040		−.553**		−.556**		−.584**	
Δ Food Away	−.329**		−.247		−.323*		−.316*		−.308*	
Model 3		.035		.225**		.038		.049		.026
Δ CPI	−.579**		−.047		−.554**		−.557**		−.585**	
Δ Food Away	−.351**		−.192		−.345**		−.341**		−.327*	
Δ Population	−.189		.477**		−.196		−.222		−.162	
Model 4		.035		.104*		.038		.047		.019
Δ CPI	−.603**		−.011		−.580**		−.585**		−.603**	
Δ Food Away	−.343**		−.204		−.337**		−.332**		−.321*	
Δ Population	−.208		.514**		−.216		−.244*		−.176	
ΔUnemployment	.190		−.326**		.199		.220		.141	

* $p < .05$.
** $p < .01$.

Table 1. Results of the Reynolds, Rahman, and Balinbin predictive sales study (2013, 321)

Similarly, Kim and Upneja completed a variation to this study where they utilized the Compustat database produced by Standard and Poor's Institutional Market Services to "examine key financial distress factors for publicly traded U.S. restaurants for the period from 1988 to 2010" (2013, 354). The dataset contains publicly listed restaurant companies from North American Industry Classification System's food services and drinking places. There are "826 observations, 42 observations from 21 companies were categorized as financially distressed, and the other 784 observations from 121 companies were sorted into non-distressed firms according to their Zmijewski scores" (Kim and Upneja 2013, 357).

Kim and Upneja developed their predictive models with Decision Trees and AdaBoosted Decision Trees. The splitting points for both trees and their prediction accuracy results (2013, 358) are in Table 2.

The significant variables for DT financial distress prediction.

	Variables	Splitter's level	Splitter's cutoff value	
Entire model	Debt-to-equity ratio	1st	>	2.962320
	Current ratio	2nd	≤	0.551677
	Net profit margin	3rd	≤	0.012932
	Account receivable turnover	4th	>	108.347418
	Current ratio	5th	≤	1.287666
Full-service model	Debt-to-equity ratio	1st	>	2.892596
	Net profit margin	2nd	≤	0.012212
	Debt-to-equity ratio	3rd	≤	4.862768[a]
	Growth in owners' equity	4th	≤	−0.404893
	Debt-to-equity ratio	3rd	>	4.862768[a]
	Account receivable turnover	4th	>	71.562753
Limited-service model	Debt-to-equity ratio	1st	>	2.408019
	Debt-to-equity ratio	2nd	>	5.395452
	Debt-to-equity ratio	2nd	≤	5.395452[a]
	Return on common equity	3rd	≤	0.160802

[a] Indicates the variables which have any direct relationship with neither distress nor non-distress.

Prediction accuracy of the proposed models.

Model		Non-distress (%)	Distress (%)	Total (%)
DT	Entire model	85.29	66.67	96.7312
	Full-service model	98.52	61.54	96.9984
	Limited-service model	93.60	56.25	90.4255
AdaBoosted DT	Entire model	98.98	73.81	97.6998
	Full-service model	99.18	73.08	98.1043
	Limited-service model	97.09	50.00	93.0851

Table 2. Kim and Upneja's splitting points and prediction accuracy results (2013, 358)

Miller conducted a study where he utilized "the autoregressive integrated moving average (ARIMA) methodology to develop forecasts for three time series of monthly archival trucking prices obtained from two public sources—the Bureau of Labor Statistics (BLS) and Truckstop.com" (2018, 130). The author chose the ARIMA methodology because it "provides a useful framework for understanding the evolution of motor carrier rates because (i) the method has a substantial degree of flexibility and (ii) the theoretical meaning of ARIMA parameters map well to dynamic forces expected to affect trucking prices" (Miller 2018, 130).

The coefficients of variation of the root-mean-square deviations for all the models are 0.007, 0.040, and 0.048. The model's equation is as follows, and the graph in Figure 2 illustrates the model's effectiveness.

$$\Delta Rate_t = \alpha_0 + \alpha_1 \Delta Rate_{t-1} + e_t + \beta_1 e_{t-1}$$

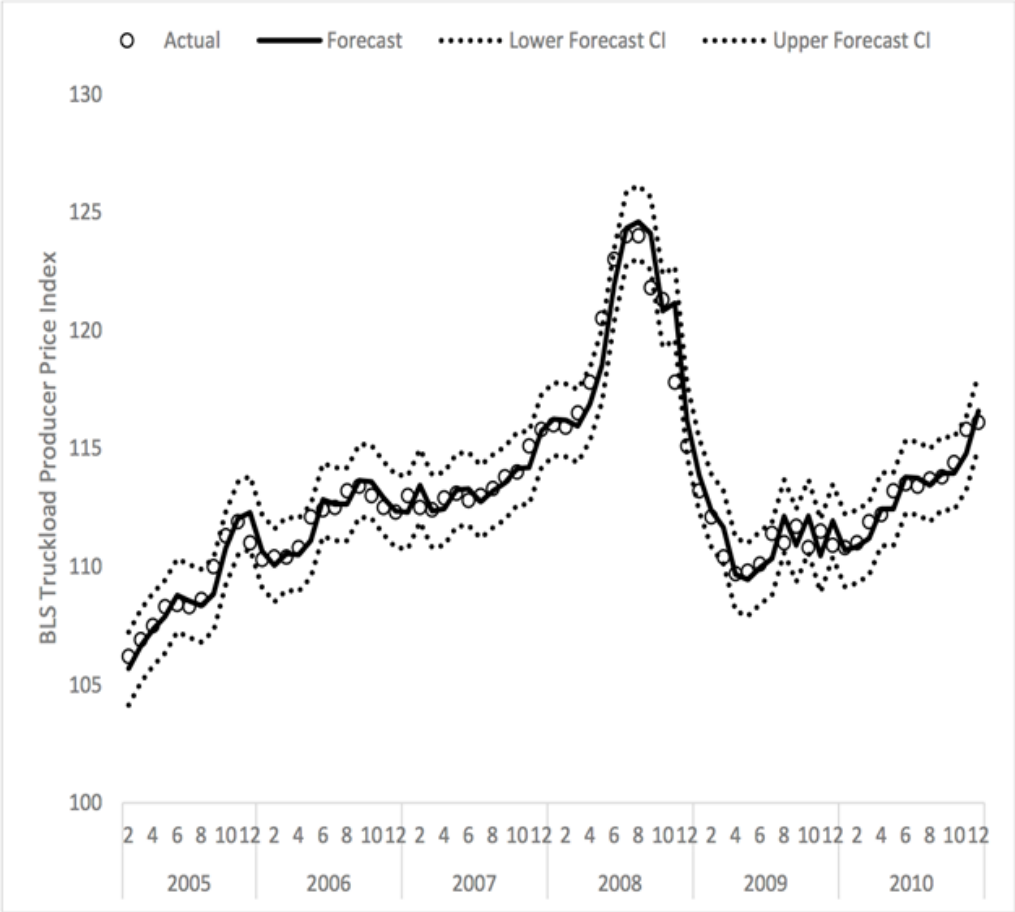


Figure 2. ARIMA example

Stepwise regression "involves relating many independent variables to a dependent variable" (Cheng et al. 2012, 619). "This method involves trying out many independent variables and including them in the regression model if they are statistically significant. If they are not statistically significant they are removed" (Cheng et al. 2012, 620). Cheng et al. (2012, 620) utilized stepwise multivariate regression and SPSS to predict the PM_{10} emissions for China's Handan and Tangshan. The independent variables of this study include the coal consumption and the annual operating cost of exhaust gas control devices for the production and supply of electric power and heating power (EP&HP) sector, the coal consumption and electricity consumption for the production of construction materials (PCM) sector, the coal consumption and other solid fuel consumption for metallurgical industry (MI) sector, and the coal

consumption for chemical processing (CP) sector. The equation, Correlation Coefficient, and Mean Error for each predicted category are as follows.

$$E_{\text{industry},i} = C_i + \sum_j a_{ij} A_{ij}$$

$$E_{\text{non-industry}} = C_{\text{non-industry}} + \sum_k b_k B_k$$

Correlation coefficient (R) and mean error (MER) of the regression model for each industrial and non-industrial sector in Handan.

Category	R	MER
Industrial sector		
EP&HP	0.997	34.0%
PCM	0.873	30.4%
MI		
Annual production value of above 1 billion Yuan	0.972	19.4%
Annual production value of below 1 billion Yuan	0.795	46.8%
CP	0.683	99.5%
Non-industrial sector	0.953	39.8%

Additional work by Jason Brownlee indicates a viable usage of long short-term memory (LSTM) neural networks for predictive sales power (Brownlee 2018, 21). The process of an LSTM is highly similar to that of a recurrent neural network (RNN) but allows for the capture of weight changes through time, improving the model's ability to learn (Brownlee 2018, 48). This process is detailed in Figure 3 and through the tutorial and subsequent discussion by Michael Phi. He discusses the notable contributions of Dr. Juergen Schmidhuber to the field of data science and as the credited father of LSTMs (Phi 2018, 1).

Yu and colleagues detail the usefulness of LSTMs in several case studies. They determine that LSTM modeling types can be helpful in sales forecasting across various business applications, including supply chain predictions and total sales (Yu et al. 2018, 15). Dai and Huang further illustrate the viability of LSTM models in predicting sales, particularly with additional model tuning that can allow for increased efficiency in modeling and better predictive performance (Dai and Huang 2021, 12015).

The predictive power of LSTM models is a quality needed for the forecasting goals of MarginEdge's sales prediction tool. From these earlier works on decision tree modeling and LSTMs, this project intends to apply various previously proven methodologies to identify which novel combination is best suited for the problem at hand.

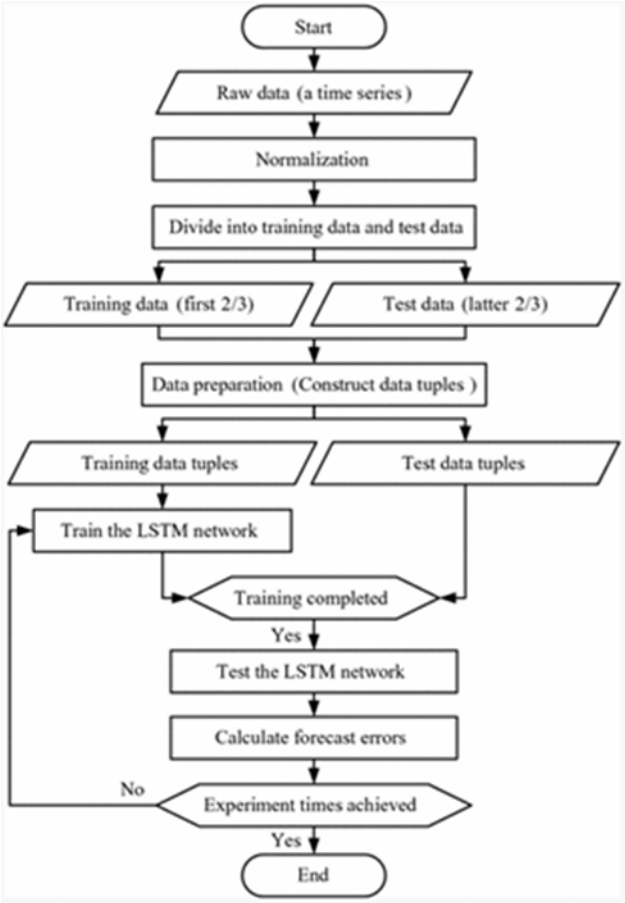


Figure 3. LSTM process as detailed by Jason Brownlee (2018)

Data

Exploratory Data Analysis

MarginEdge provided a summary of their customers' sales data in Virginia, Maryland, and the District of Columbia (DC) from early 2011 to mid-April 2021 for analysis. The dataset represented 380 restaurants and contained 385,177 observations and thirteen features. The features included a unique identifier for the restaurant, sales date, POS system used, the restaurant's zip code, an indicator for whether the restaurant is full service, and the total sales for the day, along with a breakout of seven sales subcategories. All features contained nonnull values. Refer to Appendix A for the associated data dictionary.

The initial exploratory analysis demonstrated the variability to which restaurant sales increase or decrease. Figure 4 illustrates how the average sales of all restaurants on a given day have evolved. The sales oscillate daily with decreasing peaks at the end of 2016, 2017, 2018, and 2019. There is a significant dip at the start of 2020's second quarter, which we can attribute to the Covid-19 pandemic, with March 16 as a good characterization of 2020's lowest sales point. Figure 5 expands in more detail the pandemic timeframe from March 2020 to mid-April 2021,

including other dips that can attribute to additional spikes and closures as related to Covid-19 surges.

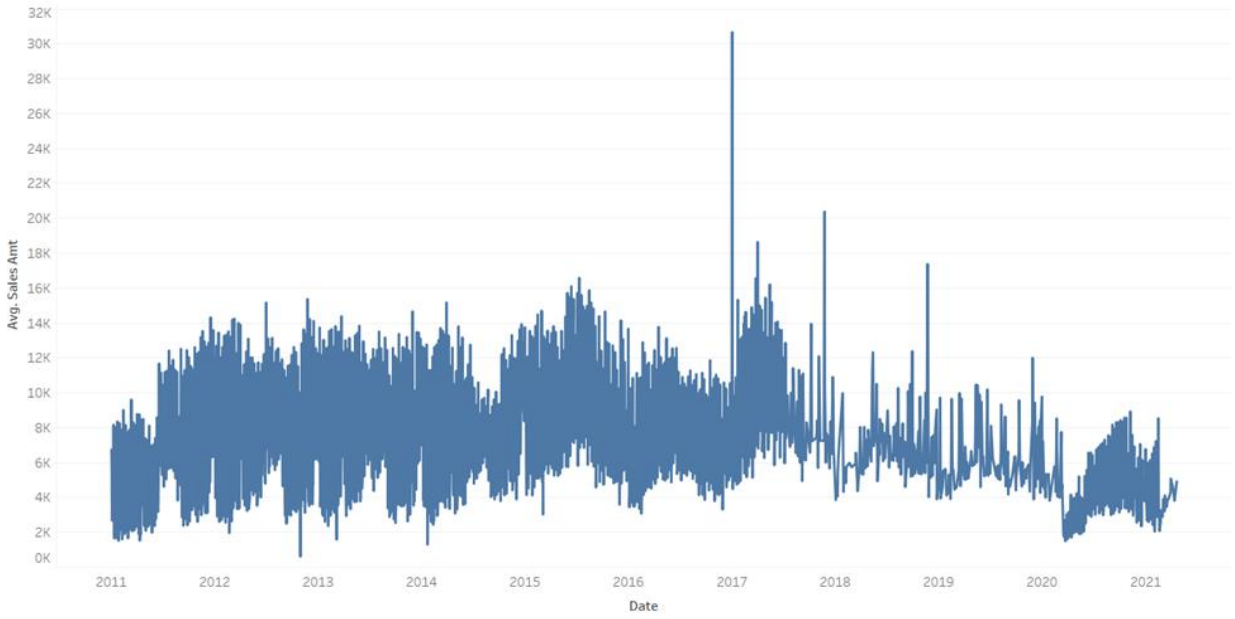


Figure 4. Average sales of all restaurants over time

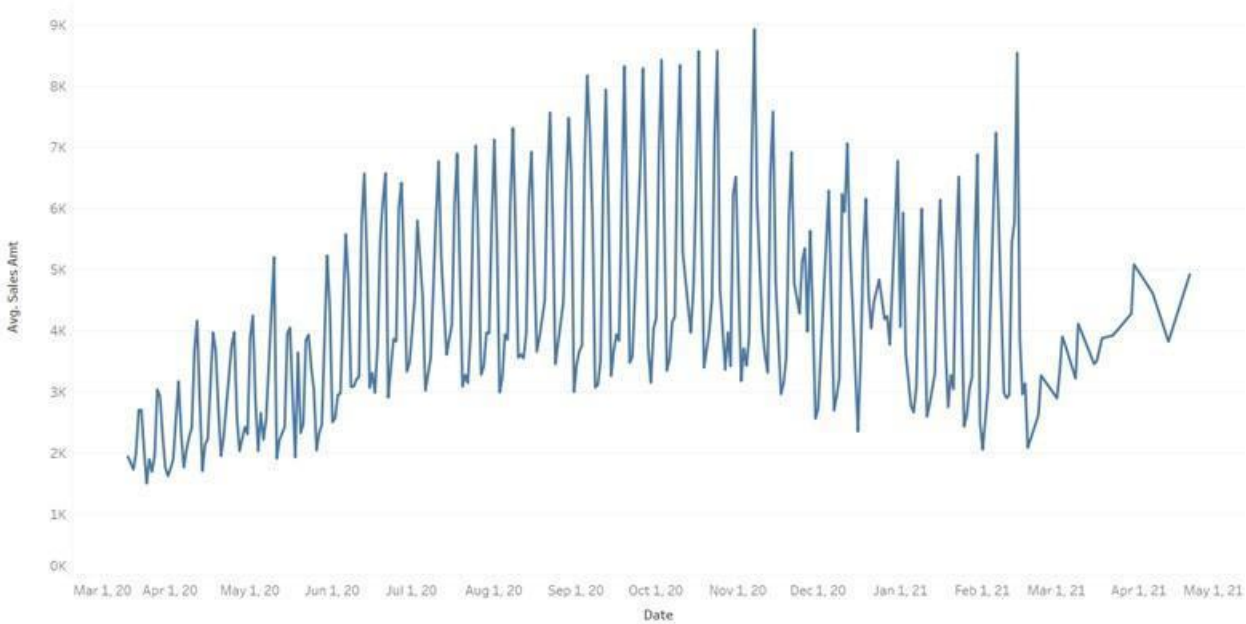


Figure 5. Overall restaurant sales from the onset of the pandemic to mid-April 2021

Figure 6 illustrates the dichotomy of the average sales differential between full-service restaurants and non-full-service restaurants. Restaurants that provide full service have higher revenues than restaurants that do not have that offering. When developing our models, we considered restaurants may act inherently differently by the service types, especially during economic changes and the pandemic.

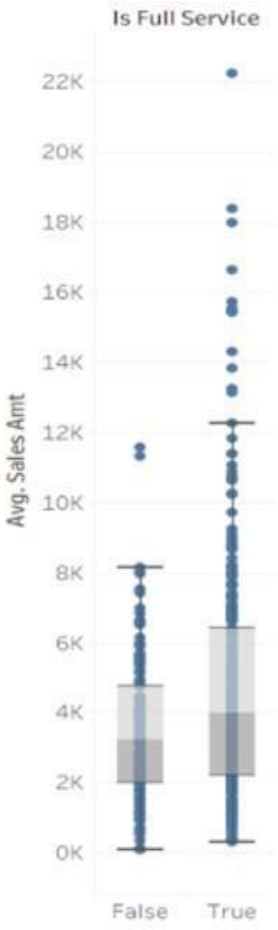


Figure 6. Boxplot Service Type by Sales

We noted that sales were affected by many U.S. holidays. Refer to Appendix B for a breakout by Christmas, New Year's Day, Valentine's Day, Independence Day, and Halloween. Also, there was significant variability in the day of the week as shown in Appendix C, indicating it may have strong predictive power. However, the sales pattern by day differed by restaurant which may be problematic in using a generalized model.

We concluded the following from our exploratory data analysis (EDA):

- Overall sales dipped significantly due to the pandemic but are starting to pick up. Therefore, we wanted to normalize the pandemic data separately from the pre-pandemic data to give a sense of what future lockdowns or recovery periods could hold for restaurants.
- Holidays can be somewhat predictive, but that predictive quality is dependent upon the holiday in question.
- Demographic data may be useful indicator of which areas were most affected by the pandemic and are rebounding faster.
- Service mode is an excellent indicator of restaurant sales. Restaurants that are in full service bring significantly higher revenue than restaurants that are not.
- Day of the week is an excellent indicator of restaurant sales. All restaurants we observed profit more on certain days than they do on other days.
- The features provided in the original dataset may be sufficient for restaurants with consistent sales, but they largely ignore external factors.

To compensate for lack of features pertaining to external circumstances, we created 160 additional features from external sources and by engineering existing features. In addition to historical sales features, we captured restaurant reviews and price ranges, Covid-19 seating capacity allowed on the sales date, zip code demographics, and indicators of U.S. holidays. The graphs in Appendix D highlight the differences we identified by zip code. Refer to Appendix E for the data dictionary for the dataset used for modeling.

Data Cleaning

The combined sales subcategories differed from the total sales for the restaurants in numerous instances. We removed the observations associated with the twenty-five restaurants had differences greater than $10,000 (n = 17,389). For restaurants with a total variance less than or equal to $10,000, we retained their total sales value and applied the variance to the food subcategory. Additionally, we identified some negative values in the sales total and subcategories. However, these could be valid transactions due to discounts or refunds, so we opted to retain those observations. We did not believe that the POS system added value as an explanatory value, so we eliminated the POS feature and combined the day's total sales by restaurant, which reduced the number of observations by 264 to 367,524.

An observation count by year revealed that MarginEdge had had substantial growth since 2011. We concurred the observations before 2016 were too minimal to be included in the overall population. This exclusion further reduced the dataset by 8,414 observations. Further, ee assumed that missing dates for the restaurants were due to closure for holidays or similar circumstances. We added 30,482 observations for restaurants with missing dates between the first and last dates MarginEdge had received sales and set the observations' sales to $0.

We created additional features related to historical sales. To ensure these features were not underinflated, we removed 133,621 observations where MarginEdge did not have at least 56 weeks of data before the given sales date. We removed observations for two restaurants that had extreme poor data quality. The remaining number of observations (n) included for modeling was 253,378 and represented 353 restaurants. Table 3 illustrates the changes in the number of observations.

	Observations (n)	
Description of Change to Dataset	**Change**	**Cumulative**
MarginEdge original dataset	385,177	385,177
Excluded: Restaurants where difference between total sales and subcategories is greater than $10,000	(17,389)	367,788
Combined: POS records by day and restaurant	(264)	367,524
Added: Missing days from 2016+ for restaurants with $0 sales *(presumed to be closed)*	30,482	398,006
Excluded: Transactions prior to 2016	(8,414)	389,592
Excluded: Transactions where MarginEdge did not have historical sales data for 56 weeks + prior to sale date	(133,621)	255,971
Excluded: Two restaurants with poor data quality	(2,593)	253,378
Ending: Observations included for modeling		**253,378**

Table 3. Changes to observations included in dataset

Research Design & Methods

It is preferred to fit the predictive model on all restaurants collectively for ease of implementation. Therefore, we explored this method using stepwise regression and multiple linear regression. If either method is successful compared to fit the model, we would use all restaurants to fit the model. Otherwise, only the restaurant specific data would be used to fit corresponding model for four randomly selected models.

We were also interested in determining the impact of Covid-19 in the training data. Therefore, we also evaluated 01/01/16-04/14/21 training data and post Covid-19 training data from 04/15/20-04/14/21 (the "training" datasets). We applied the following techniques to predict the sales by day.

- Simple linear regression (LR)
- Ordinary least squares regression (OLSR)
- Weighted simple linear regression (Weighed LR)
- Weighted ordinary least squares regression (Weighted OLSR)
- Long Short-Term Memory (LSTM)
- Stepwise Regression (SR)
- Multiple Linear Regression (MLR)
- Decision Tree Regression (DT)
- Various Autoregression Integrated Moving Average (ARIMA) approaches

To optimize the model's fit, we created multiple models for each modeling technique that is discussing further in the Analysis & Interpretation section. The model with the highest R^2 against the training dataset that was considered optimal for each modeling method. The root mean squared error was another method that was considered. Next, the top fit for each modeling technique was validated against a test set that consisted of daily sales from 04/15/21-04/21/21. The most important metric in evaluating the models was the average absolute variance as a percent of daily sales. We also considered the variance with a 5% tolerance allowed and the R^2. The model with the lowest average absolute variance as a percent of daily sales is deemed optimal for implementation.

Results

Trained on All Restaurants

The multiple linear regression model that was fit on data from all restaurants since 2016 appeared to fit the restaurant data well according to the results in the following table. It only has an average absolute daily variance of 4.5% and 1.7% with a 5% percent tolerance allowed on the test data set. 99% of the variability in the sales was explainable from the variables selected.

		Method				
Date	Actual	Stepwise Regression		Multiple Linear Regression		n
04/15/21	$ 1,618,590	$ 1,765,535	$ 1,672,290	$ 1,723,136	$ 1,716,482	309
04/16/21	$ 2,510,359	$ 2,457,411	$ 2,180,366	$ 2,463,185	$ 2,446,014	309
04/17/21	$ 3,117,941	$ 2,799,192	$ 2,384,494	$ 3,065,517	$ 3,002,387	309
04/18/21	$ 2,318,026	$ 2,070,714	$ 1,926,785	$ 2,274,084	$ 2,250,540	308
04/19/21	$ 1,012,211	$ 1,340,804	$ 1,372,575	$ 1,133,101	$ 1,180,594	303
04/20/21	$ 1,358,185	$ 1,479,022	$ 1,463,651	$ 1,324,664	$ 1,341,384	299
04/21/21	$ 1,298,166	$ 1,656,655	$ 1,622,127	$ 1,489,718	$ 1,489,895	294
Total	$13,233,478	$ 13,569,333	$ 12,622,288	$13,473,405	$13,427,296	2,131
Training Data:	Restaurants	*353*	*353*	*353*	*353*	
	Time frame	*2016+*	*Post Covid*	*2016+*	*Post Covid*	
Metric						
R-Squared		0.965	0.964	0.990	**0.991**	
Root Mean Squared Error		2,843,039	**2,730,186**	2,864,418	2,853,381	
Ave Abs. Daily Variance %		11.9%	17.4%	**4.5%**	5.5%	
Daily Var % w/ Tolerance		7.4%	12.6%	**1.7%**	2.0%	

When we reviewed our four randomly selected restaurants, it became apparent that the model trained on all restaurants' data fit well on the restaurant industry. However, it did not fit well for individual restaurants as illustrated in the following table. The average absolute daily variance ranged from 17% to 113.9% d depending on the restaurant while R^2 ranged from 0.402 to 0.98. Since our objective to accurately predict restaurant sales by restaurant, it was clear that the aggregated approach was not appropriate. As outlined in the Research Design & Methods section, we trained the remainder of the models on the data for restaurant of interest since the aggregated data fit failed to meet our needs.

	Method			
	Stepwise Regression		Multiple Linear Regression	
Training Data Time frame	*2016+*	*Post Covid*	*2016+*	*Post Covid*
Ave Abs. Daily Variance %				
Restaurant A	132.6%	91.0%	113.9%	24.7%
Restaurant B	47.9%	49.6%	17.0%	14.6%
Restaurant C	28.1%	31.2%	19.3%	18.9%
Restaurant D	32.7%	37.1%	25.2%	23.1%
R-Squared				
Restaurant A	0.191	0.179	0.402	0.443
Restaurant B	0.889	0.884	0.980	0.982
Restaurant C	0.772	0.743	0.850	0.864
Restaurant D	0.698	0.699	0.786	0.883

Trained on Restaurant of Interest

The result by metric by restaurant is in Appendix F. The optimal result each is highlighted in blue. For Restaurants B and D, the Hold-Winters ARIMA model predicted sales within an average daily variance of only 9.5% and 5.1%, respectively. The multiple linear regression model trained on the post Covid data had the best results (24.7%) for Restaurant A while the decision tree regression model trained on the previous 45 days produced the strongest results for Restaurant C (14.5%).

Analysis and Interpretation

For illustrative purposes, we referred to the results for Restaurant D throughout the Analysis and Interpretation section.

Ordinary Least Squares Regression & Simple Linear Regression

For both regression types, we wanted to explore a simple model that predicts the restaurant's past 45 days of data and an ensemble model that predicts the total sales from a weighted sum of each anticipated subcategory sale. To build an accurate ensemble model using weighted sum, we determined how each subcategory relates to the total sales. We created a multivariate regression model that predicts the sales amount with the sales amount of each subcategory to see if this is a viable option. The following figures illustrates Restaurant D's model fit, and the high R^2 value indicates the model may provide strong predictions for this restaurant. The zero p-values of all coefficients show that sales from each subcategory directly contributed to the total sales.

Dep. Variable:	sales_amt	**R-squared (uncentered):**	0.996
Model:	OLS	**Adj. R-squared (uncentered):**	0.996
Method:	Least Squares	**F-statistic:**	4.020e+06
Date:	Tue, 27 Apr 2021	**Prob (F-statistic):**	0.00
Time:	16:21:30	**Log-Likelihood:**	-9.5281e+05
No. Observations:	126030	**AIC:**	1.906e+06
Df Residuals:	126023	**BIC:**	1.906e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sales_amt_food	1.0000	0.000	2780.324	0.000	0.999	1.001
sales_amt_liquor	1.0000	0.002	530.913	0.000	0.996	1.004
sales_amt_wine	1.0220	0.003	326.110	0.000	1.016	1.028
sales_amt_nabeverages	0.9961	0.013	79.323	0.000	0.971	1.021
sales_amt_other	1.0008	0.000	2138.884	0.000	1.000	1.002
sales_amt_beer	0.9959	0.003	358.287	0.000	0.990	1.001
sales_amt_retail	1.0021	0.009	117.388	0.000	0.985	1.019

Date	Actual	OLSR	Abs. Variance
04/15/21	$ 2,449	$ 4,903	$ 2,455
04/16/21	$ 4,851	$ 4,830	$ 21
04/17/21	$ 9,986	$ 4,854	$ 5,131
04/18/21	$ 7,492	$ 4,879	$ 2,613
04/19/21	$ 3,211	$ 4,842	$ 1,631
04/20/21	$ 2,566	$ 4,866	$ 2,301
04/21/21	$ 3,109	$ 4,891	$ 1,782
Total	$ 33,663	$ 34,065	$ 403

Metric	OLSR
R^2	0.070
Root Mean Squared Error	7,346
Ave Abs. Daily Variance %	47.3%
Daily Var % w/ Tolerance	43.0%

However, despite the optimistic training results, OLSR and linear regression models developed we not effective by industry standards. As the following table shows, the OLSR was the best model, but there was significant daily variance, especially on 04/17/21. The mean absolute percentage error of 47.3% shows that the prediction of each day is on average 47.3% higher or lower than the actual value. The model over generalized its predictions and failed to discriminate based on the day of the week. For a business that is running on many perishable resources and needed to schedule resources, this is not viable.

Bidirectional LSTM

The Bidirectional Long Short-Term Memory (LSTM) model, a cousin of recurrent neural networks, was selected to evaluate the applicability of a high computational power neural network to the business need at hand. Leaders in the LSTM field indicate the routine usage of rectified linear activation function (ReLU) as the activator, Adam as the optimizer function and mean squared error as the loss metric for training and testing, and these parameters were utilized in this experiment for consistency with no other hyper-parameters being added. The same 6-week date range was chosen for an appropriate comparison to the other models tested and 30 epochs were run for resource management. Because of the relatively short time frame of training data, a bidirectional LSTM was chosen to maximize model performance over a short training span.

For Restaurant D, the bidirectional LSTM model had an R^2 value of 0.34, a root mean squared error of 6,539.37 and an absolute daily variance of 35.21%. It performed poorly against the test data for all restaurants evaluated. These results illustrate that despite having quite a bit of computational power, the bidirectional LSTM was no more effective at predicting the daily sales than less complex model types that were evaluated. The lack of high caliber performance from the bidirectional LSTM is likely due to overfitting of the model; it struggled to generalize well across large and small fluctuations in actual sales amounts. It is likely the residual fluctuations of Covid-related changes played a part in the overfitting as well as normal restaurant sale total variations across days of the week.

Stepwise Regression & Multiple Linear Regression

We applied the forward-selection rule to the stepwise regression model. The explanatory variables were chosen by an automatic procedure that begins with no explanatory variables. Then, the procedure iteratively adds the most statistically significant variable until there are not any remaining statistically significant variables remaining. The final explanatory variables set for Restaurant D contained 42 variables. The following regression formula could be used to predict Restaurant D's sales for a given day.

$\hat{y}$ = -2257.32 + 1164.57*[Open Status] + 1482.73*[Seating Capacity Allowed] + 62.72[quarter] + -308.45*[is Monday] + 688.47*[is Friday] + 2086.19*[is Saturday] + 1119.18*[is Sunday] + -10114.44*[is Christmas] + -2598.98*[is Thanksgiving] + 4221*[is Columbus Day] + -6902.09*[is Easter] + -1117.04*[is Father's Day] + 3346.3 *[is Good Friday] + -1916.94*[is Halloween] + 5057.46*[is Labor Day] + 5887.42*[is Martin Luther King Day] + 4967.12*[is Memorial Day] + 4370.91*[is President's Day] + 1617.52*[is Valentine's] + 2074.08*[is Veteran's Day] + 0.34*[Ave. Sales in Past 7 Days] + 0.39*[Ave. Sales on Same Past 2 Weekdays] + 0.47*[Ave. Prior Year Sales a Week Before and After]+ -6551.74*[Ave. Seating Capacity Allowed a Week Before and After in Prior Year] + -0.16*[Ave .Prior Year Sales 2 Weeks Before] + -0.38*[Ave. Prior Year Sales 4 Weeks Before] + 5825.29*[Ave. Seating Capacity Allowed 4 Weeks Before in Prior Year] + -0.01*[Ave. Prior Year Sales 3 Similar Weekdays] + 0.29[Ave. Prior Year Sales on Same Weekday]

The stepwise method provided stronger results than LR, OLSR, and LSTM because it incorporated factors related to holidays, days of the week, and seating capacity allowed. Interestingly, the demographic and yelp data were not selected as explanatory variables. This is likely because we only had a snapshot of May 2021 instead of being able to compare values over time. Also, since we were training the data at the restaurant level, these snapshot values never changed to provide predictive power. To potentially add value, we would need the data as of the sales date.

However, the top average absolute daily variable was still too high at 21.3%. We consistently under predicted sales, which could result in poor customer service and decreased revenue. Also, the complexity of the model may be problematic for implementation.

		Stepwise Regression			
Training Data Time frame		*2016+*	*Post Covid*	*2016+*	*Post Covid*
Date	Actual	Predictions		Absolute Variance	
04/15/21	$ 2,449	$ 3,468	$ 3,307	$ 1,019	$ 858
04/16/21	$ 4,851	$ 5,355	$ 5,525	$ 504	$ 674
04/17/21	$ 9,986	$ 7,630	$ 7,910	$ 2,356	$ 2,075
04/18/21	$ 7,492	$ 5,152	$ 5,023	$ 2,340	$ 2,469
04/19/21	$ 3,211	$ 2,649	$ 2,627	$ 563	$ 585
04/20/21	$ 2,566	$ 2,610	$ 2,241	$ 44	$ 325
04/21/21	$ 3,109	$ 2,773	$ 2,466	$ 336	$ 643
Total	$ 33,663	$ 29,636	$ 29,099	$ 4,027	$ 4,564
Metric					
R-Squared		0.868	0.849		
Root Mean Squared Error		7,160	7,163		
Ave Abs. Daily Variance %		21.3%	22.7%		
Daily Var % w/ Tolerance		16.5%	17.7%		

We also developed multiple linear regression model where we chose the explanatory variables through an iterative process. The variables were selected based on the training data for all restaurants for sales after 2016, and some variables were removed due to multicollinearity. The variables selected were used to fit on only the data for the restaurant of interest. As a result, each restaurant would have unique coefficients, but the same variables would be required for the model. The final model performed yielded inconsistent results by restaurant compared to the stepwise regression model. We believe this is because we attempted to standardize the variables selected in the MLR process, whereas the stepwise regression model selected the variables based on the specific restaurant's data.

Decision Tree Regression

Decision Trees, by nature, allow restaurant owners who are not very familiar with technology to experience the power of predictive analytics. Unlike other models, the research team can generate visualizations of the decision tree model that restaurant owners can use to determine their predictions. Appendix G provides some illustrative examples of decision tree models that predict based on the day of the week. We evaluated decisions tress based on the day of the week for post Covid data and data within the past 45 days. We also developed a decision tree on post Covid data based on the month and day of the week in hopes of capturing seasonality.

The decision trees generally performed well on the training data, but the results were inconsistent against the test data. It was our top performer for Restaurant C and ARIMA was a close second. If we integrated seating capacity allow, we may have more success.

ARIMA

Training Data:		2016 +	Past 6 Weeks	Past 6 Weeks	Past 5 Weeks excl. Spring Break
Date	Actual	Holt-Winters	Holt-Winters	ARIMA (Std)	Holt-Winters
04/15/21	$ 2,449	$ 2,997	$ 3,628	$ 3,448	$ 2,881
04/16/21	$ 4,851	$ 4,996	$ 6,721	$ 6,038	$ 5,389
04/17/21	$ 9,986	$ 9,569	$ 10,732	$ 8,887	$ 9,851
04/18/21	$ 7,492	$ 7,664	$ 8,064	$ 7,158	$ 7,561
04/19/21	$ 3,211	$ 3,643	$ 3,617	$ 3,350	$ 3,111
04/20/21	$ 2,566	$ 2,574	$ 2,957	$ 2,982	$ 2,615
04/21/21	$ 3,109	$ 3,108	$ 3,581	$ 3,744	$ 2,902
Total	$ 33,663	$ 34,552	$ 39,300	$ 35,607	$ 34,310

Metric	Holt-Winters	Holt-Winters	ARIMA (Std)	Holt-Winters
R-Squared	**0.992**	0.966	0.959	0.990
Root Mean Squared Error	7,799	8,325	**7,788**	7,823
Ave Abs. Daily Variance %	5.1%	16.7%	14.3%	**4.5%**
Daily Var % w/ Tolerance	**2.1%**	27.7%	20.5%	32.1%

The Autoregressive Integrated Moving Average (ARIMA) model was used using several date ranges, but the most accurate required a relatively short time range of 6 weeks. This model provided excellent accuracy with a .96 R^2 and an absolute daily variance of 14%. While this was one of the more robust models we used, further analysis and experimentation showed that the Holt-Winters (HW) adaptation to the ARIMA model proved much more successful. The HW model allows the inclusion of a predictor variable to better weight the model on seasonality, and by providing a 7-day cycle (the calendar week), we attained superior results. As you can see from the table on the left, the HW model with a complete set of 5 years of sales data (going back to 2016) produced fantastic results with an R^2 of .99 and variance of only 5%!

Conclusions

The goal of this project was to identify a model and approach to predicting the total sales any restaurant could expect in the coming days and weeks. Through the evaluation of a variety of modeling types, predictive success was found with several different approaches and a clear viable option identified for business-scale investment and rollout. The findings of this investigation indicate that the complexity and variability of day over day sales paired with the contextual fluctuations of the current post-pandemic climate are extremely impactful factors for predictive quality across all model types tested.

High predictive success was found with the Holt Winters ARIMA, or Autoregressive Integrated Moving Average, which utilized seasonality weighting to account for dramatic swings in total sales day to day and week to week. It produced the most consistent results by restaurant. For Restaurant D, the Holt Winters ARIMA model produced an R^2 value of 0.99, indicating extremely high alignment to the actual sales reported by the restaurants evaluated in this project. These results supply MarginEdge with a viable solution to invest resources into and expand as a solution for their customers in the future. The ability for restaurants to better plan their supply chain and resourcing needs in accordance with actual vs. predicted sales can help

these businesses that have struggled so much over the past year recoup what was lost and grow for the future.

Future Work

This project serves as Phase 1 (R&D) of the product's development. MarginEdge's technical staff will carry out Phase 2 (Beta) and Phase 3 (Rollout). The details are as follows.

- Phase 2 (Beta): MarginEdge's development team will build into the core platform the pipeline for thousands of predictions (one per restaurant) to be created nightly. We will enable this functionality only in select restaurants to test. MarginEdge hopes to provide a sensitivity level around predictions for restaurant preferences. One example of this preference is overpredicting sales vs. underpredicting sales. MarginEdge's team will need to monitor forecasts across the entire client base of several thousand restaurants to look for outliers and their causes and adjustments for societal changes like with Covid-19 dining restrictions.

- Phase 3 (Rollout): Once this new predictive feature is proven reliable across a broad swath of restaurant types and companies, MarginEdge intends to deploy it full scale across its entire client base and use it as a bedrock system for additional predictive exploration.

Appendix A: Original Data Dictionary

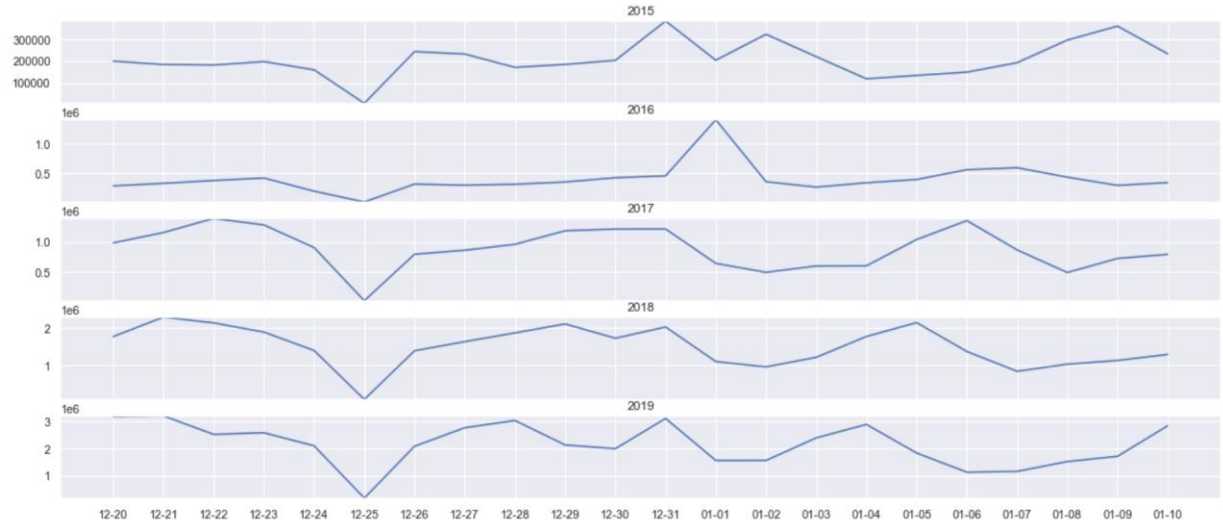
Data Dictionary for Dataset provided by MarginEdge.

Feature	Variable Type	Description
restaurant_unit_id	nominal	Unique restaurant ID by location
date	ordinal	Sales date (YYYY-MM-DD)
pos	nominal	POS system; there may be many POS systems associated with a single restaurant on a given day
is_full_service	dichotomous	1 = Full service restaurant with wait staff; 0 = restaurant that does not have wait staff
zip	nominal	Zip code in which the restaurant is located
sales_amt_food	continuous	Subcategory of sales that represents the amount of food purchased in dollars
sales_amt_liquor	continuous	Subcategory of sales that represents the amount of liquor purchased in dollars
sales_amt_wine	continuous	Subcategory of sales that represents the amount of wine purchased in dollars
sales_amt_nabeverages	continuous	Subcategory of sales that represents the amount of nonalcoholic beverages purchased in dollars; for restaurants that do not serve alcohol, these values will typically be included in food
sales_amt_other	continuous	Subcategory of sales that represents the any purchases, discounts, reimbursements, or other atypical transactions that are not represented in the other sales categories.
sales_amt_beer	continuous	Subcategory of sales that represents the amount of beer purchased in dollars
sales_amt_retail	continuous	Subcategory of sales that represents the amount of retail items purchased in dollars; such items may include, but are limited to, clothing, kitchenware, and magnets
sales_amt	continuous	Total sales

Appendix B: Sales by Holiday Period

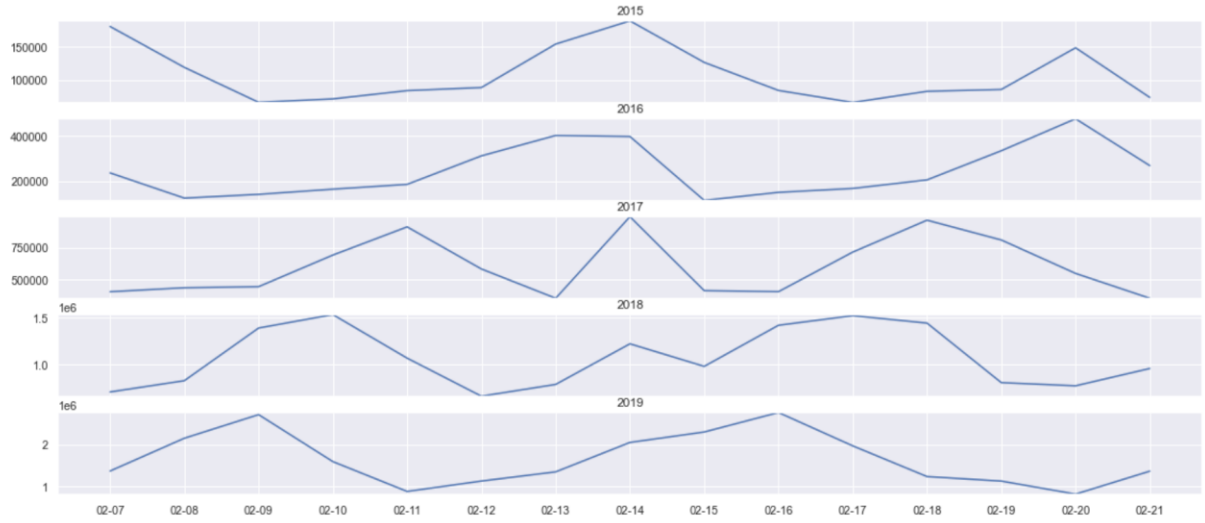
Holiday sales patterns emerge year over year that can be valuable insights for our predictive models. Sales decreased significantly before Christmas but picked up again after the holiday. This trend is likely due to closures and folks staying home with loved ones before venturing out again to celebrate the New Year.

Christmas/New Year's



People had a strong urge to dine at restaurants before, during, and after Valentine's Day, indicating that the surrounding timeframe and the holiday fell on the weekend can play essential factors for this holiday.

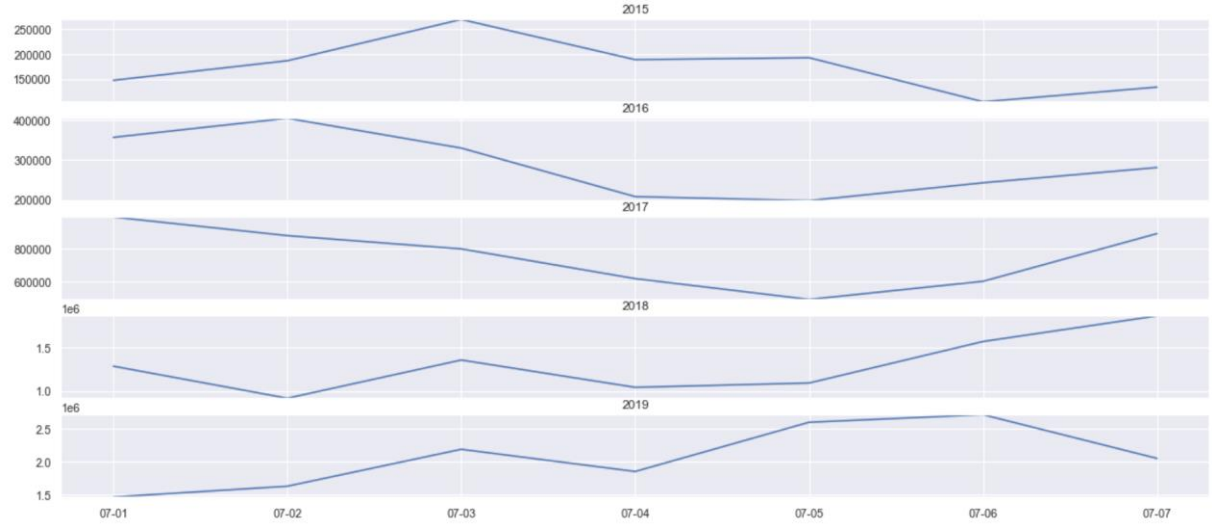
Valentine's Day



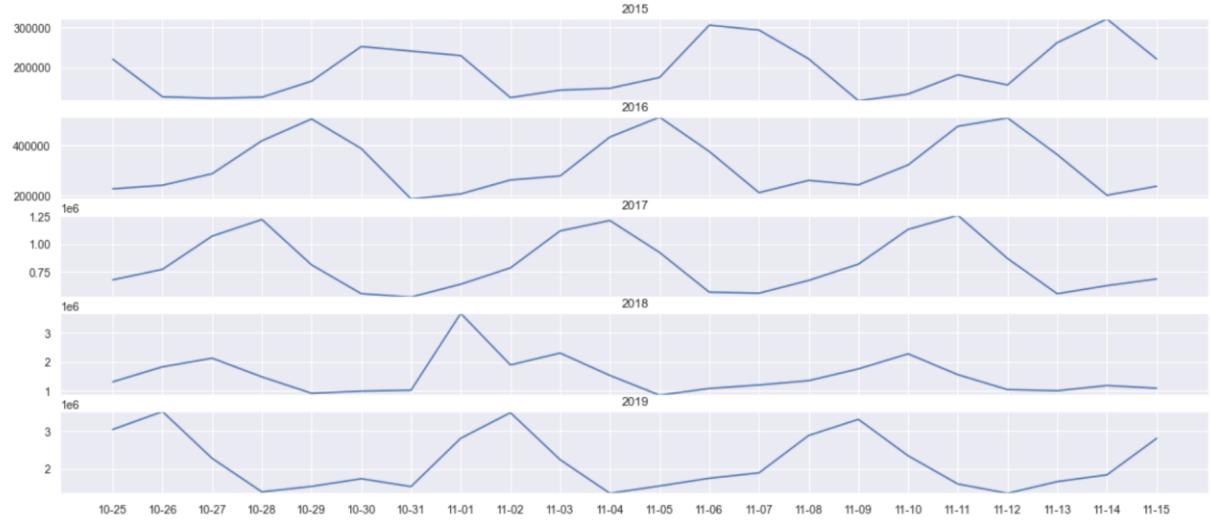
Sales by Holiday Period continued.

It is a lack of interest in restaurant purchases on or near the Fourth of July. We can attribute this frequency of fairs and barbecues going on at that time. The frequency of restaurant purchases on Halloween is mixed through the years and does not easily lend itself to be an excellent predictive holiday.

Fourth of July

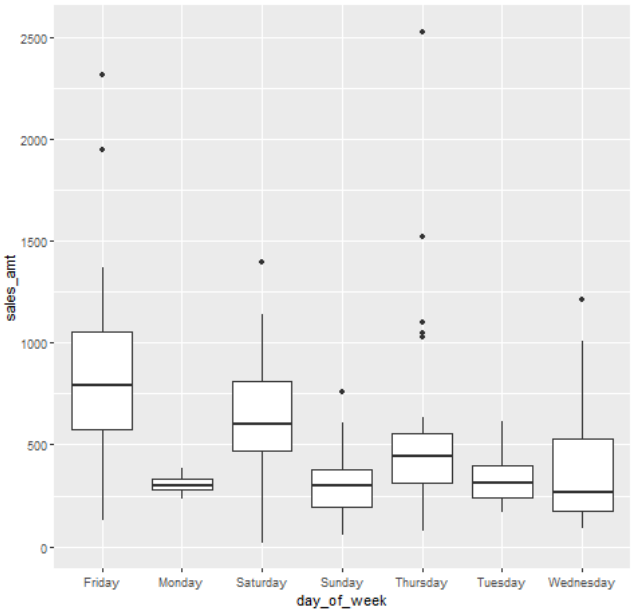


Halloween

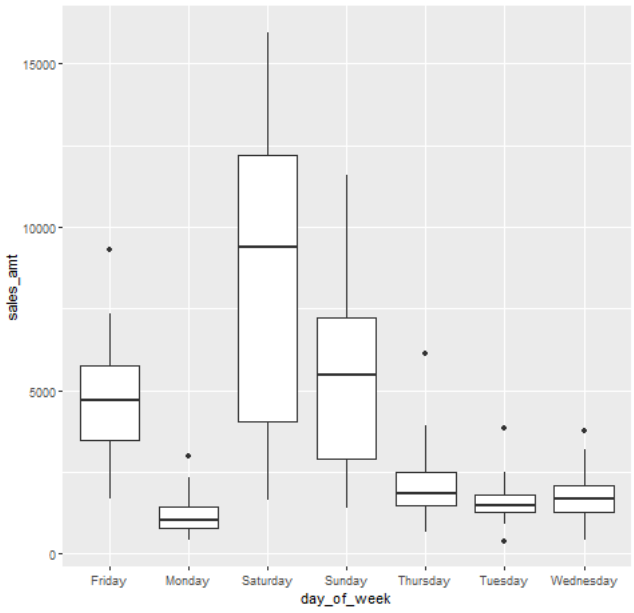


Appendix C: Boxplot of Sales by Day of the Week by Restaurant

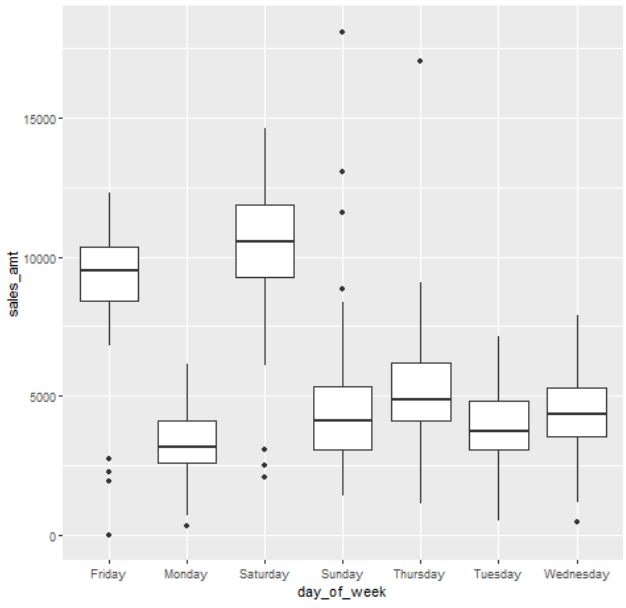
Restaurant A



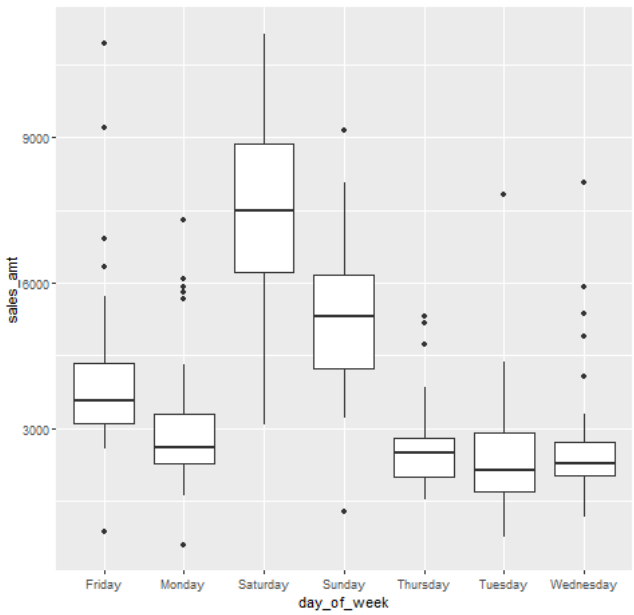
Restaurant B



Restaurant C

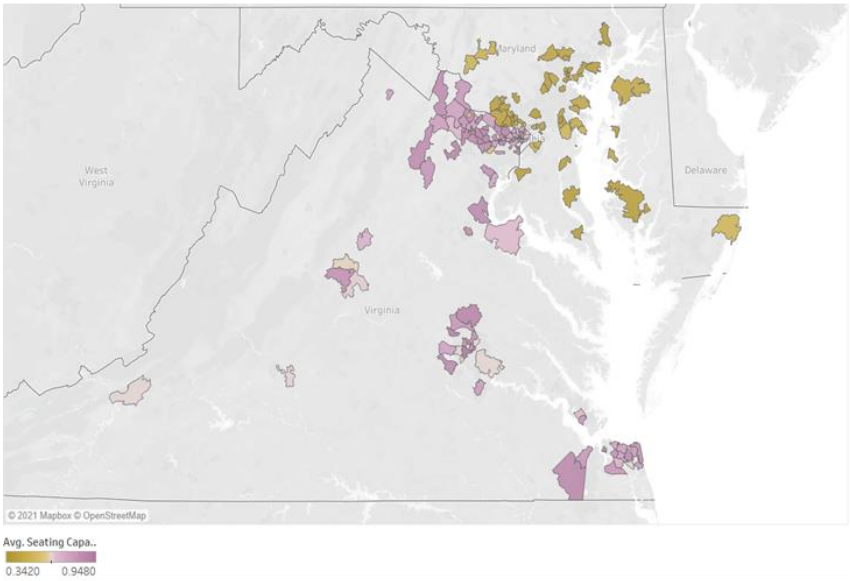


Restaurant D



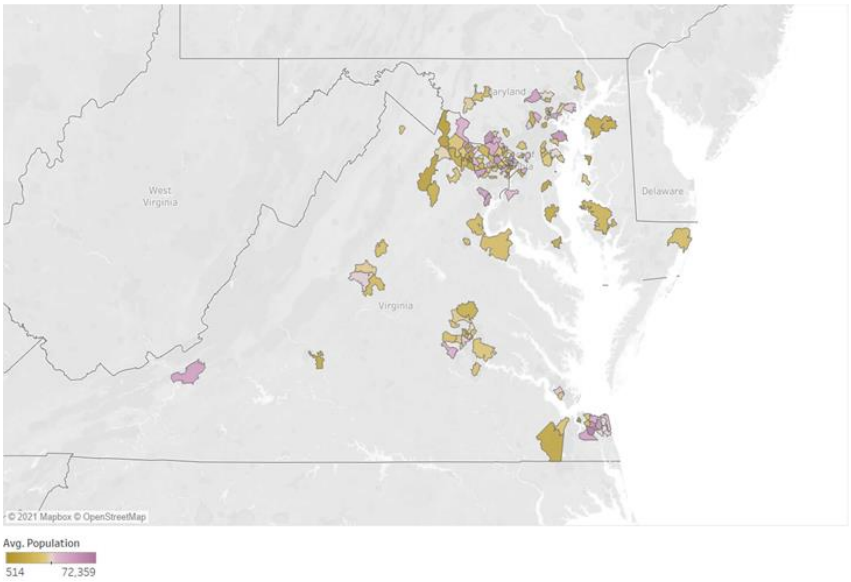
Appendix D: Variable of Interest by Zip Code

Allowable Seating Capacity by Zip Code after the Pandemic



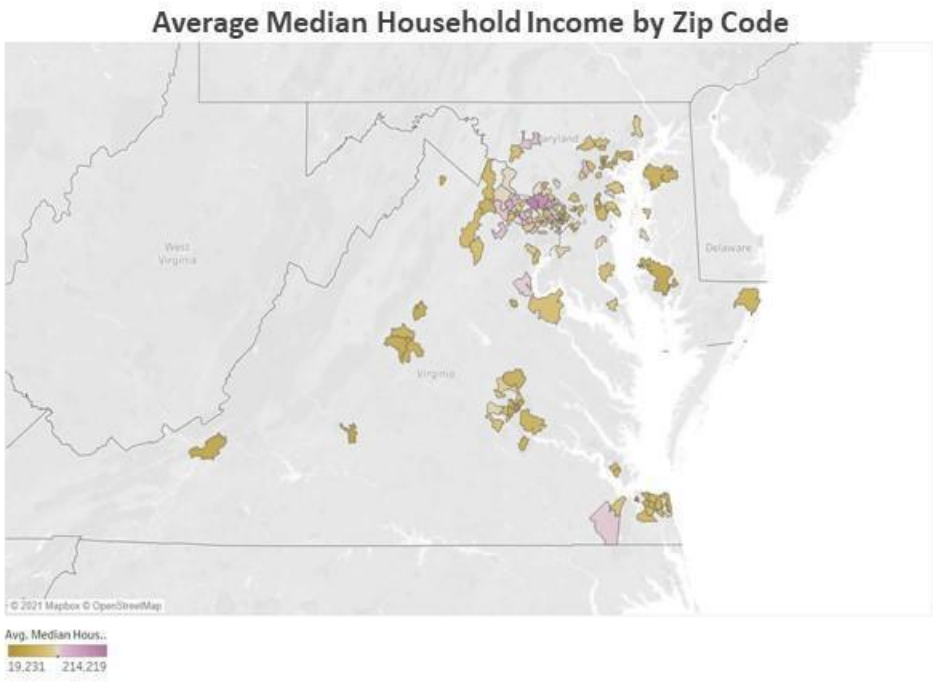
Based on the allowable average seating capacity of MarginEdge's clients grouped by zip code, there appear to be significant variations in the seating allowed, which is likely to affect restaurant sales by region.

Average Population by Zip Code

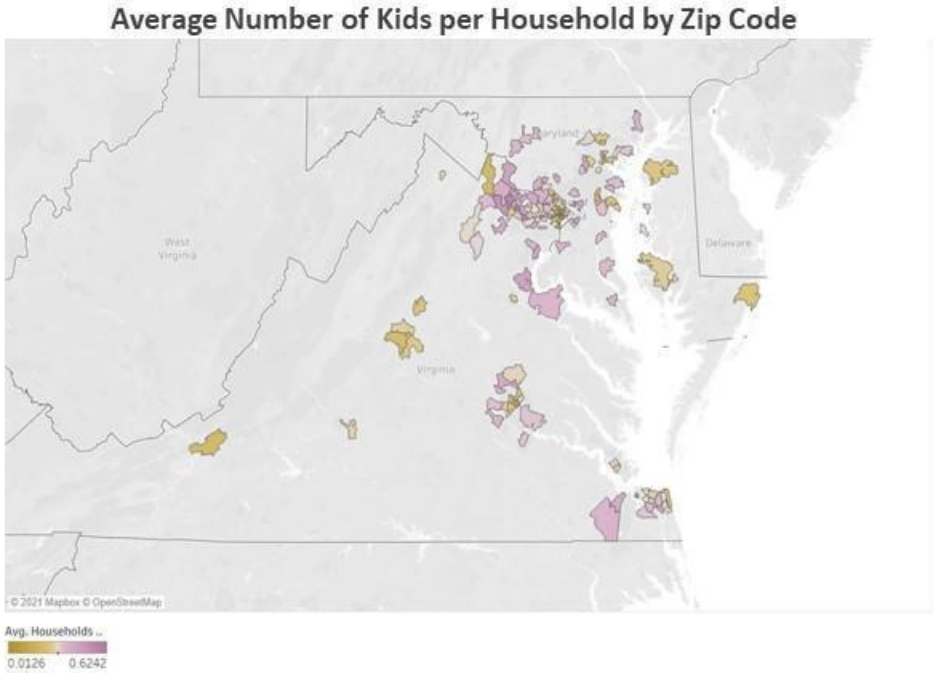


The average population by zip code of MarginEdge's clients indicates that the zip codes on the outer portion of the sample region are generally more populated.

Variable of Interest by Zip Code continued.



The average median household income by zip code illustrates that areas near Maryland and the District of Columbia have a much higher median income and presumably more expendable income.



The average number of kids per household may provide helpful information about the average sales by region.

Appendix E: Complete Data Dictionary

Complete data dictionary for the dataset used in the analyses.

Feature	Source Type	Description
Response Variable		
sales_amt	0	Total sales
Sales Date Specific Information		
OpenStatus	1	1 = Presumed open (sales); 0 = Presumed closed (no sales for day, but MarginEdge has sales before and after date)
date	0	Sales date (YYYY-MM-DD)
SeatingCapacity	3	Estimated seating capacity allowed per local or state restrictions (1 = 100% seating capacity allowed)
day_of_mo	1	Numeric days of the month
day_of_week_nbr	1	Numeric representation of the day of the week (0 = Sunday,…,0 = Saturday)
day_of_week	1	Day of the week
week	1	Number of week in the year
month	1	Numeric representation of the month (1 = January,…,12 = December)
quarter	1	Numeric representation of the quarter(1 = 1st quarter,…,4 = 4th quarter)
year_half	1	1 = January to June; 2 July to December
year	1	Sales year
is_Sunday	1	1 = Sunday; 0 = Any other weekday
is_Monday	1	1 = Monday; 0 = Any other weekday
is_Tuesday	1	1 = Tuesday; 0 = Any other weekday
is_Wednesday	1	1 = Wednesday; 0 = Any other weekday
is_Thursday	1	1 = Thursday; 0 = Any other weekday
is_Friday	1	1 = Friday; 0 = Any other weekday
is_Saturday	1	1 = Saturday; 0 = Any other weekday
is_weekend	4	1 = Saturday or Sunday; 0 = Any other weekday
is_christmas	4	1 = Christmas holiday; 0 = Any other day
is_columbus_day	4	1 = Columbus day holiday; 0 = Any other day
is_easter_sunday	4	1 = Easter Sunday holiday; 0 = Any other day
is_fathers_day	4	1 = Father's Day holiday; 0 = Any other day
is_good_friday	4	1 = Good Friday holiday; 0 = Any other day
is_halloween	4	1 = Halloween holiday; 0 = Any other day
is_independence_day	4	1 = Christmas holiday; 0 = Any other day
is_labor_day	4	1 = Labor Day holiday; 0 = Any other day
is_martin_luther_king_day	4	1 = Martin Luther King Day Jr holiday; 0 = Any other day
is_memorial_day	4	1 = Memorial Day holiday; 0 = Any other day

Source Types:

0 = Original dataset
1 = Engineered from original dataset
2 = https://www.unitedstateszipcodes.org/
3 = https://www.huschblackwell.com/
4 = https://www.timeanddate.com/holidays/us/
5 = https://www.yelp.com/

Feature	Source Type	Description
Sales Date Specific Information		
is_mothers_day	4	1 = Mother's Day holiday; 0 = Any other day
is_national_donut_day	4	1 = National Donut Day holiday; 0 = Any other day
is_new_years_day	4	1 = New Year's Day holiday; 0 = Any other day
is_presidents_day	4	1 = President's Day holiday; 0 = Any other day
is_thanksgiving_day	4	1 = Thanksgiving Day holiday; 0 = Any other day
is_valentines_day	4	1 = Valentine's Day holiday; 0 = Any other day
is_veterans_day	4	1 = Veteran's Day holiday; 0 = Any other day
is_holiday	4	1 = Holiday identified; 0 = Any other day
Restaurant's historical sales and days open relative to the observation's sale date.		
H_07daysSales	1	Total sales for 7 days prior to the record's sale date
H_07daysOpen	1	Number of days with sales for 7 days prior to the record's sale date
H_07daysAveSales	1	For 7 days prior to the record's sale date, average sales on "open days"; Null if closed all days
H_07daysSeatgCap	3	For 7 days prior to the record's sale date, average seating capacity allowed
H_14daysSales	1	Total sales for 14 days prior to the record's sale date
H_14daysOpen	1	Number of days with sales for 14 days prior to the record's sale date
H_14daysAveSales	1	For 14 days prior to the record's sale date, average sales on "open days"; Null if closed all days
H_14daysSeatgCap	3	For 14 days prior to the record's sale date, average seating capacity allowed
H_2WeekdaySales	1	Total sales for 7th day and 14th day prior to the record's sale date (same weekday)
H_2WeekdayOpen	1	Number of days with sales associated with 7th day and 14th day prior to the record's sale date (same weekday)
H_2WeekdayAveSales	1	Average sales on "open days" for the 7th day and 14th day prior to the record's sale date; Null if closed all days
H_2WeekdaySeatgCap	3	For 14 days prior to the record's sale date, average seating capacity allowed
H_PrevYr_4WksPrior_Sales	1	Total sales between 52-56 weeks before current date
H_PrevYr_4WksPrior_Open	1	Number of days with sales between 52-56 weeks before current date
H_PrevYr_4WksPrior_AveSales	1	For 52-56 weeks prior to the record's sale date, average sales on "open days"; Null if closed all days
H_PrevYr_4WksPrior_SeatgCap	3	For 52-56 weeks prior to the record's sale date, average seating capacity allowed
H_PrevYr_2WksPrior_Sales	1	Total sales between 52-56 weeks before current date
H_PrevYr_2WksPrior_Open	1	Number of days with sales between 52-54 weeks before current date

Source Types:

0 = Original dataset

1 = Engineered from original dataset

2 = https://www.unitedstateszipcodes.org/

3 = https://www.huschblackwell.com/

4 = https://www.timeanddate.com/holidays/us/

5 = https://www.yelp.com/

Feature	Source Type	Description
Restaurant's historical sales and days open relative to the observation's sale date.		
H_PrevYr_2WksPrior_AveSales	1	For 52-54 weeks prior to the record's sale date, average sales on "open days"; Null if closed all days
H_PrevYr_2WksPrior_SeatgCap	3	For 52-54 weeks prior to the record's sale date, average seating capacity allowed
H_PrevYr_1wkBothWays_Sales	1	Total sales between 51-53 weeks before current date (max 15 days)
H_PrevYr_1wkBothWays_Open	1	Number of days with sales between 51-53 weeks before current date (max 15 days)
H_PrevYr_1wkBothWays_AveSales	1	For 51-53 weeks prior to the record's sale date, average sales on "open days"; Null if closed all days
H_PrevYr_1wkBothWays_SeatgCap	3	For 51-53 weeks prior to the record's sale date, average seating capacity allowed
H_PrevYr_SameWkDay_Sales	1	Total sales 52 weeks (day only) before current date
H_PrevYr_SameWkDay_Open	1	Number of days with sales 52 weeks (day only) before current date
H_PrevYr_SameWkDay_AveSales	1	Total sales 52 weeks (day only) before current date
H_PrevYr_SameWkDay_SeatgCap	3	Seating capacity allowed 52 weeks (day only) before current date
H_PrevYr_3WkDays_Sales	1	Total sales exactly 51, 52, and 53 weeks before current date (max 3 days)
H_PrevYr_3WkDays_Open	1	Number of days exactly 51, 52, and 53 weeks before current date (max 3 days)
H_PrevYr_3WkDays_AveSales	1	Average daily sales for exactly 51, 52, and 53 weeks before current date (max 3 days)
H_PrevYr_3WkDays_SeatgCap	3	Average seating capacity allowed exactly 51, 52, and 53 weeks before current date (max 3 days)
Restaurant Level Information		
restaurant_unit_id	0	Unique restaurant ID by location
is_full_service	0	1 = Full service restaurant with wait staff; 0 = restaurant that does not have wait staff
zip_code	0	Zip code in which the restaurant is located
State	2	State in which the restaurant is located
First_Date	1	First sales date reported
Last_Date	1	Last sales date reported
Total_Days_Open	1	For entire original dataset, number of days with sales transactions
Ave_Sales_Per_Day	1	Total Sales divided by Days Reported
Perc_Food	1	For entire original dataset, total food sales (as defined in this dictionary) divided by total sales
Perc_Alcohol	1	For entire original dataset, combined liquor, beer, and wine sales divided by total sales
Perc_Retail	1	For entire original dataset, retail sales divided by total sales
Perc_Other	1	For entire original dataset, "other" sales divided by total sales

Source Types:

0 = Original dataset
1 = Engineered from original dataset
2 = https://www.unitedstateszipcodes.org/
3 = https://www.huschblackwell.com/
4 = https://www.timeanddate.com/holidays/us/
5 = https://www.yelp.com/

Feature	Source Type	Description
Restaurant Level Information		
Offers_NoAlcohol	1	1 = no alcohol offered; 0 = alcohol offered (determined by alcohol sales)
Offers_BeerWineOnly	1	1 = beer and wine offered but liquor is not; 0 = liquor offered in addition to beer and wine (determined by alcohol sales)
Offers_Liquor	1	1 =liquor is offered; 0 = liquor is not offered (determined by alcohol sales)
Offers_Alcohol	1	1 = alcohol offered ; 0 = no alcohol offered (determined by alcohol sales)
Yelp Data		
yelp_PriceGroup	5	Price grouping based on priceRange (1, 2, 3, or 4); from cheapest to most expensive
priceRange	5	Values: Under $10,Inexpensive, $11-30, US$11-30, Moderate, Not Found, $31-60, Above $61, or NULL
Rating	5	yelp rating ranging from 1 to 5 (lowest to highest); contains NULLs when no rating has been applied
NbrOfReviews	5	Number of reviews given on yelp
CuisineType	5	Primary cuisine type listed on yelp (52 distinct values)
YelpURL	5	Restaurant specific URL from which data was scraped
YelpCity	5	City per yelp
YelpState	5	State per yelp
Zip Code's Stats and Demographics		
Population	2	Zip code's population
Population_Density	2	Zip code's people per square mile
Housing_Units	2	Zip code's housing units
Median_Home_Value	2	Zip code's median home value
Land_Area	2	Zip code's land area per square mile
Water_Area	2	Zip code's water area per square mile
Occupied_Housing_Units	2	Zip code's number of occupied housing units
Median_Household_Income	2	Zip code's median household income
Median_Age	2	Zip code's median age
Male_Median_Age	2	Zip code's median age of males
Female_Median_Age	2	Zip code's median age of females
Zip Code's Gender Stats		
Male	2	Percent of males in zip code
Female	2	Percent of females in zip code

Source Types:

0 = Original dataset

1 = Engineered from original dataset

2 = https://www.unitedstateszipcodes.org/

3 = https://www.huschblackwell.com/

4 = https://www.timeanddate.com/holidays/us/

5 = https://www.yelp.com/

Feature	Source Type	Description
Zip Code's Race Stats		
White	2	Percent of people who are White in zip code
Black_or_African_American	2	Percent of people who are Black or African American in zip code
American_Indian_Or_Alaskan_Native	2	Percent of people who are American Indian Or Alaskan Native in zip code
Asian	2	Percent of people who are Asian in zip code
Native_Hawaiian_Other_Pacific_Islan	2	Percent of people who are Native Hawaiian Other Pacific Islander in zip code
Other_Race	2	Percent of people who are Other Race in zip code
Two_Or_More_Races	2	Percent of people who are Two Or More Races in zip code
Zip Code's Family vs. Singles Stats		
Husband_Wife_Family_Households	2	Percent of households that are "Husband, Wife, Family" in zip code
Single_Guardian	2	Percent of households that are "Single and Guardian" in zip code
Singles	2	Percent of households that are "Singles" in zip code
Singles_With_Roommate	2	Percent of households that are "Singles With Roommate "in zip code
Zip Code's Households with Kids		
Households_without_Kids	2	Percent of Households without Kids in zip code
Households_with_Kids	2	Percent of Households with Kids in zip code
Zip Code's Housing Types		
In_Occupied_Housing_Units	2	Percent of housing types that are "In Occupied Housing Units" in zip code
Correctional_Facility_For_Adults	2	Percent of housing types that are "Correctional Facility For Adults" in zip code
Juvenile_Facilities	2	Percent of housing types that are "Juvenile Facilities in zip code
Nursing_Facilities	2	Percent of housing types that are "Nursing Facilities" in zip code
Other_Institutional	2	Percent of housing types that are "Other Institutional" in zip code
College_Student_Housing	2	Percent of housing types that are "College Student Housing" in zip code
Military_Quarters	2	Percent of housing types that are "Military Quarters" in zip code
Other_Noninstitutional	2	Percent of housing types that are "Other Noninstitutional" in zip code
Zip Code's Housing Occupancy		
Owned_Households_With_A_Mortgage	2	Percent of homes that are "Owned Households With A Mortgage" in zip code
Owned_Households_Free_Clear	2	Percent of homes that are "Owned Households Free Clear" in zip code
Renter_Occupied_Households	2	Percent of homes that are "Renter Occupied Households" in zip code
Households_Vacant	2	Percent of homes that are "Households Vacant" in zip code

Source Types:

0 = Original dataset

1 = Engineered from original dataset

2 = https://www.unitedstateszipcodes.org/

3 = https://www.huschblackwell.com/

4 = https://www.timeanddate.com/holidays/us/

5 = https://www.yelp.com/

Feature	Source Type	Description
Zip Code's Vacancy Reasons		
For_Rent	2	Percent of vacancy reasons that are "For Rent" in zip code
Rented_Unoccupied	2	Percent of vacancy reasons that are "Rented Unoccupied" in zip code
For_Sale_Only	2	Percent of vacancy reasons that are "For Sale Only" in zip code
Sold_Unoccupied	2	Percent of vacancy reasons that are "Sold Unoccupied" in zip code
For_Season_Recreational_Or_Occasional_Use	2	Percent of vacancy reasons that are "For Season Recreational Or Occasional Use" in zip code
For_Migrant_Workers	2	Percent of vacancy reasons that are "For Migrant Workers" in zip code
Vacant_For_Other_Reasons	2	Percent of vacancy reasons that are "Vacant For Other Reasons" in zip code
Zip Code's Rental Properties Number of Rooms		
Studio_Apartment	2	Percent of rental properties that are Studio Apartment in zip code
_1_Bedroom	2	Percent of rental properties that with 1 bedroom in zip code
_2_Bedroom	2	Percent of rental properties that with 2 bedrooms in zip code
_3_Bedroom	2	Percent of rental properties that with 3 bedrooms in zip code
Zip Code's Employment Status		
Worked_Full_time_with_Earnings	2	Percent of employment status of "Worked Full time with Earnings" in zip code
Worked_Part_time_with_Earnings	2	Percent of employment status of "Worked Part time with Earnings" in zip code
No_Earnings	2	Percent of employment status of "No Earnings" in zip code
Zip Code's Source of Earnings		
Worked_Full_time_with_Earnings2	2	Percent of source of earnings as "Worked Full time with Earnings" in zip code
Worked_Part_time_with_Earnings2	2	Percent of source of earnings as "Worked Part time with Earnings" in zip code
No_Earnings2	2	Percent of source of earnings as "No Earnings" in zip code
Zip Code's Means Of Transportation To Work for Workers 16 and Over		
Car_truck_or_van	2	Percent of Car truck or van in zip code
Public_transportation	2	Percent of Public transportation in zip code
Taxicab	2	Percent of Taxicab in zip code
Motorcycle	2	Percent of Motorcycle in zip code
Bicycle_Walked_or_Other_Means	2	Percent of Bicycle Walked or Other Means in zip code
Worked_at_Home	2	Percent of Worked at Home in zip code

Source Types:

0 = Original dataset
1 = Engineered from original dataset
2 = https://www.unitedstateszipcodes.org/
3 = https://www.huschblackwell.com/
4 = https://www.timeanddate.com/holidays/us/
5 = https://www.yelp.com/

Feature	Source Type	Description
Zip Code's Educational Attainment For The Population 25 Years And Over		
Less_than_High_School_Diploma	2	Percent of population >= 25 with highest education level of Less than High School Diploma in zip code
High_School_Graduate	2	Percent of population >= 25 with highest education level of High School Graduate in zip code
Associate_s_degree	2	Percent of population >= 25 with highest education level of Associate's degree in zip code
Bachelor_s_degree	2	Percent of population >= 25 with highest education level of Bachelor's degree in zip code
Master_s_degree	2	Percent of population >= 25 with highest education level of Master's degree in zip code
Professional_school_degree	2	Percent of population >= 25 with highest education level of Professional school degree in zip code
Doctorate_degree	2	Percent of population >= 25 with highest education level of Doctorate degree in zip code
Zip Code's School Enrollment (Ages 3 to 17)		
Enrolled_in_Public_School	2	Percent of children ages to 17 who are Enrolled in Public School in zip code
Enrolled_in_Private_School	2	Percent of children ages to 17 who are Enrolled in Private School in zip code
Not_Enrolled_in_School	2	Percent of children ages to 17 who are Not Enrolled in School in zip code

Source Types:

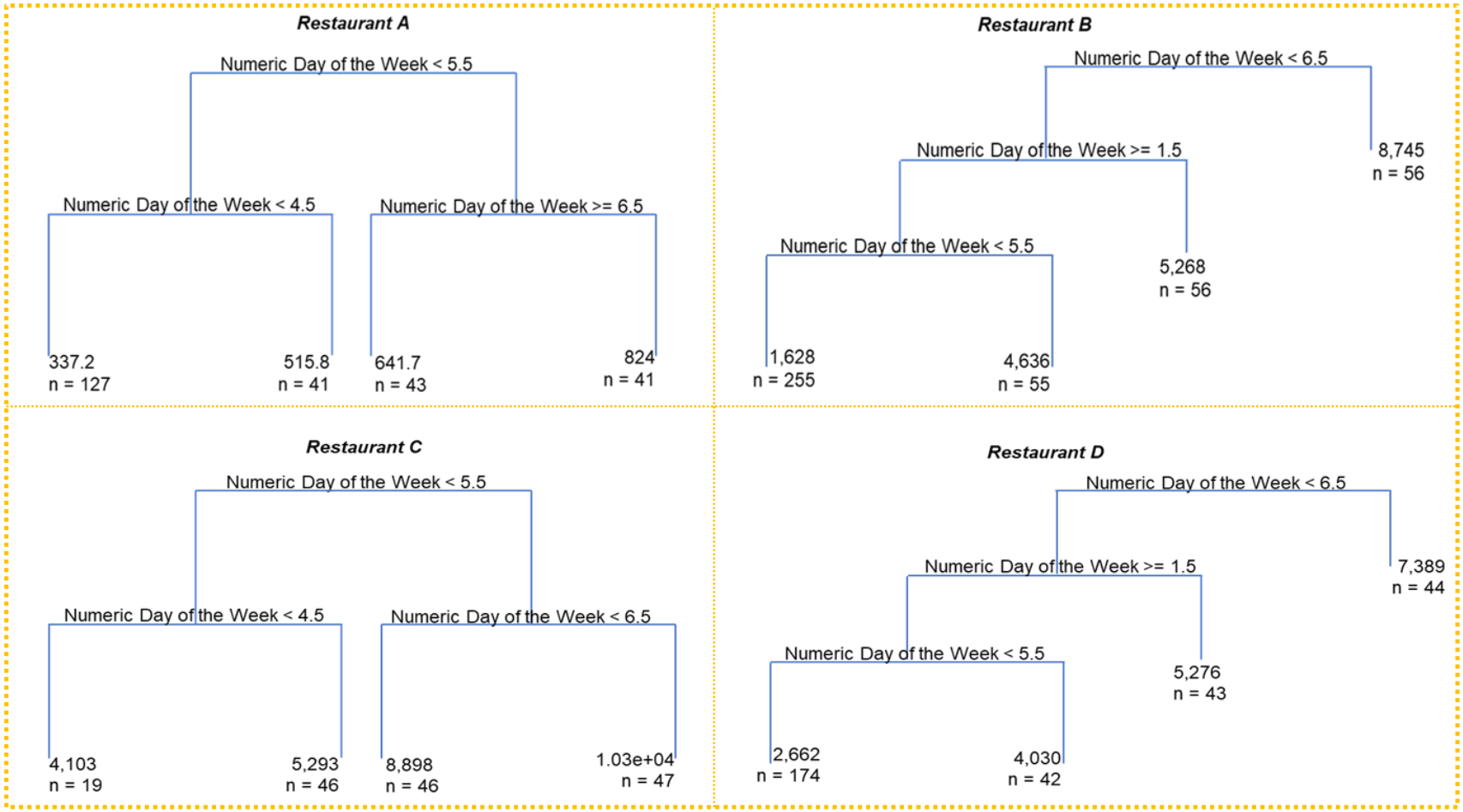
0 = Original dataset
1 = Engineered from original dataset
2 = https://www.unitedstateszipcodes.org/
3 = https://www.huschblackwell.com/
4 = https://www.timeanddate.com/holidays/us/
5 = https://www.yelp.com/

Appendix F: Results by Restaurant

	Method												
	OLSR	LR	Weighted LR	Weighted OSLR	LSTM	Stepwise Regression		Multiple Linear Regression		Decision Tree Regression			Holt-Winters ARIMA
Training Data Time frame	*Post Covid*	*Post Covid*	*Post Covid*	*Post Covid*	*Post Covid*	*2016+*	*Post Covid*	*2016+*	*Post Covid*	*Past 45 Days*	*Post Covid*	*Post Covid*	*2016+*
Ave Abs. Daily Variance %													
Restaurant A	34.7%	36.5%	469.3%	640.9%	30.1%	25.7%	25.7%	139.6%	**24.7%**	35.4%	50.6%	35.1%	31.2%
Restaurant B	67.4%	68.5%	67.9%	73.7%	56.5%	17.4%	10.5%	18.9%	10.4%	41.7%	36.5%	46.3%	**9.5%**
Restaurant C	38.4%	42.1%	35.2%	38.6%	33.9%	21.3%	21.5%	19.3%	19.1%	**14.5%**	24.4%	22.2%	16.6%
Restaurant D	47.3%	51.1%	52.8%	67.6%	35.2%	21.3%	22.7%	28.0%	21.6%	20.6%	20.6%	16.9%	**5.1%**
Ave Abs. Daily Variance % with Tolerance													
Restaurant A	29.7%	31.5%	464.3%	635.9%	25.1%	21.8%	21.8%	134.6%	**19.8%**	30.6%	45.6%	30.1%	26.3%
Restaurant B	62.4%	63.5%	62.9%	68.9%	51.5%	12.9%	6.1%	13.9%	5.4%	36.7%	31.5%	41.3%	**4.7%**
Restaurant C	33.4%	37.1%	30.4%	33.6%	29.2%	16.3%	16.5%	14.7%	14.5%	**10.2%**	19.6%	17.2%	11.8%
Restaurant D	43.0%	46.1%	47.8%	62.6%	30.4%	16.5%	17.7%	23.1%	16.6%	15.7%	15.7%	12.3%	**2.1%**
R-Squared													
Restaurant A	0.055	0.055	0.055	0.055	**0.709**	0.417	0.417	0.378	0.443	0.308	0.506	0.375	0.377
Restaurant B	0.146	0.146	0.146	0.146	0.072	0.985	0.982	0.981	0.984	0.605	0.986	0.671	**0.987**
Restaurant C	0.206	0.203	0.203	0.203	0.589	0.808	0.797	0.863	0.855	0.900	**0.907**	0.871	0.824
Restaurant D	0.070	0.070	0.063	0.063	0.344	0.868	0.849	0.785	0.875	0.983	0.983	0.956	**0.992**
Root Mean Squared Error													
Restaurant A	1,342	1,317	5,585	7,216	1,503	1,315	1,315	2,568	1,379	1,375	**1,161**	1,307	1,285
Restaurant B	8,866	9,090	9,018	10,110	8,944	9,705	9,795	9,923	9,823	9,452	8,408	**7,958**	10,405
Restaurant C	10,892	11,293	**9,912**	10,915	13,114	11,007	11,006	11,091	11,105	11,080	10,404	10,657	11,513
Restaurant D	7,346	7,642	7,770	9,014	**6,539**	7,160	7,163	8,200	7,892	6,958	6,958	7,157	7,799

Appendix G: Decision Tree Sketch for Each Selected Restaurant

The selected decision trees use all post Covid-19 sales data before the chosen date to predict sales for the restaurant selected based on the day of the week.



Bibliography

Brownlee, Jason. *How to Develop LSTM Models for Time Series Forecasting.* Self-published, Machine Learning Mastery, 2018.

Chen, Tong, Hongzhi Yin, Hongxu Chen, Lin Wu, Hao Wang, Xiaofang Zhou, and Xue Li. "TADA: Trend Alignment with Dual-Attention Multi-Task Recurrent Neural Networks for Sales Prediction." In *2018 IEEE International Conference on Data Mining (ICDM)*, 49–58. IEEE, 2018. https://doi.org/10.1109/ICDM.2018.00020.

Cheng, Shuiyuan, Ying Zhou, Jianbing Li, Jianlei Lang, and Haiyan Wang. "A New Statistical Modeling and Optimization Framework for Establishing High-Resolution PM10 Emission Inventory – I. Stepwise Regression Model Development and Application." *Atmospheric Environment* 60 (July 11, 2012): 613–22. https://doi.org/10.1016/j.atmosenv.2012.07.056.

Dai, Yun, and Jinghao Huang. "A Sales Prediction Method Based on LSTM with Hyper-Parameter Search." *Journal of Physics. Conference Series* 1756, no. 1 (2021): 12015–. https://doi.org/10.1088/1742-6596/1756/1/012015.

Kim, Soo Y., and Arun Upneja. "Predicting Restaurant Financial Distress Using Decision Tree and AdaBoosted Decision Tree Models." *Economic Modelling* 36 (2014): 354–62. https://doi.org/10.1016/j.econmod.2013.10.005.

Miller, Jason W. "ARIMA Time Series Models for Full Truckload Transportation Prices." *Forecasting* 1, no. 1 (2018): 121–34. https://doi.org/10.3390/forecast1010009.

Phi, Michael. "Illustrated Guide to LSTM's and GRU's: A Step by Step Explanation." *Towards Data Science* (blog). *Median*. September 24, 2018. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

Phi, Michael. "Illustrated Guide to LSTM's and GRU's: A Step by Step Explanation." YouTube Video, 11:17, September 19, 2018, https://www.youtube.com/watch?v=8HyCNIVRbSU&t=6s.

Reynolds, Dennis, Imran Rahman, and William Balinbin. “Econometric Modeling of the U.S. Restaurant Industry.” *International Journal of Hospitality Management* 34 (2013): 317–23. https://doi.org/10.1016/j.ijhm.2013.04.003.

Yu, Quan, Kesheng Wang, Jan Ola Strandhagen, and Yi Wang. “Application of Long Short-Term Memory Neural Network to Sales Forecasting in Retail—A Case Study.” *Lecture Notes in Electrical Engineering*, 2018, 11–17. https://doi.org/10.1007/978-981-10-5768-7_2.