Do Meat Toppings Affect Pizza Sales?

Datasci 203: Lab Report 2

Blake Bleier, Reese Carlton, Nicholas Lin, & Peter Valverde

Contents

Introduction

Restaurants belong in a competitive industry where success is determined by a delicate balance of food quality, service, and financial judgement. To understand financial decisions restaurants need to carefully manage costs, employ strategic pricing strategies, and maximize sales. With technology continuing to grow and impact our everyday lives, it can be beneficial for non-data driven industries, such as restaurants, to involve data into their decision making. One possibility to do so is to leverage order data to enhance sale strategies and overall performance. By analyzing order histories, restaurants might be able to unravel valuable insights into customer preferences, ordering patterns, and popular menu items. Understanding customer behaviors through this data can allow restaurants to tailor their offerings, optimize menus, and strategically price items. For pizza restaurants, analyzing the correlation between pizza ingredients and sales can reveal valuable insights into customer preferences and optimize their offerings optimally. Understanding the impact of pricing alongside ingredient combinations can guide decisions on pricing strategies or seasonal promotions, ultimately boosting sales and customer satisfaction.

The goal of this study is to estimate what types of ingredients can significantly influence pizza sales, by using artificial data for a pizza restaurant. By running a set of regression models, we attempt to estimate the values that pizza ingredients have on pizza sales.

Data, Methodology, and Modeling Decisions

The data in this study is a data set made for Plato's Pizza, a fictitious pizza restaurant based in New Jersey. It was made publicly available by a group called Maven Analytics. It includes about a year's worth of 48620 pizza orders, where each row shows the details about the order such as date and time, number of pizzas, type of pizzas, size, quantity, price, and ingredients.

To organize and operationalize sales for the data, individual pizza orders were aggregated into monthly sales per type of pizza. Then, pizza ingredients were split into their respective categories of meat, sauce, cheese, and vegetables. For each pizza, the counts of each ingredient category was then tallied. Using this table and the ingredients list of each type of pizza, we totaled up how many of each ingredient category each pizza contains (ie. a bbq chicken pizza contains one meat, four vegetables, one alternative sauce, and zero alternative cheeses).

To account for time series nature of the data (month over month sales), we decided to both one hot encode an indicator for each month and include an integer representation of month (1-12) for our regression models. These variables allow for the model to account for both individually more or less popular months as well as general trends in popularity (sales) throughout the year. After organizing the data in this manner, we then used a random probability sampling method to split the data into either an exploratory data set (with 25% probability) for exploratory data analysis and feature engineering and a modeling data set (with 75% probability) used for our regression analysis. In our final data sets, each data point represents a single type of pizza in a particular month and its associated sales and ingredient information. The final exploration and model building data sets had 102 and 283 data points respectively.

No observations were intentionally removed from the dataset. The analysis was able to be conducted on the complete dataset available for Plato's Pizza, and no observations were excluded due to missing values or other criteria.

After creating the data as described above, we began our exploratory data analysis on the exploratory split of the data set. This yielded a number of additional findings for feature engineering. Lead by our EDA, we created indicators for whether or not a pizza is vegetarian, contains any alternative cheese (not just mozzarella), or contains an alternative sauce (not just red sauce). We also observed that the relationship between number of meats on a pizza and sales appeared non-linear and thus we elected to include a squared term on number of meats in order to capture the apparent curvature in the relationship.

The only covariates that were removed was the month of January as the months were one hot encoded from month_01 to month_12.

Exploratory Data Analysis

The top box plot in *Figure 1* shows that, on average, non-vegetarian pizzas have higher Total Sales than vegetarian pizzas. The broader distribution and presence of outliers in the box plot suggest a wider variability in Total Sales for non-vegetarian pizzas, prompting further investigation into influencing factors for each category.

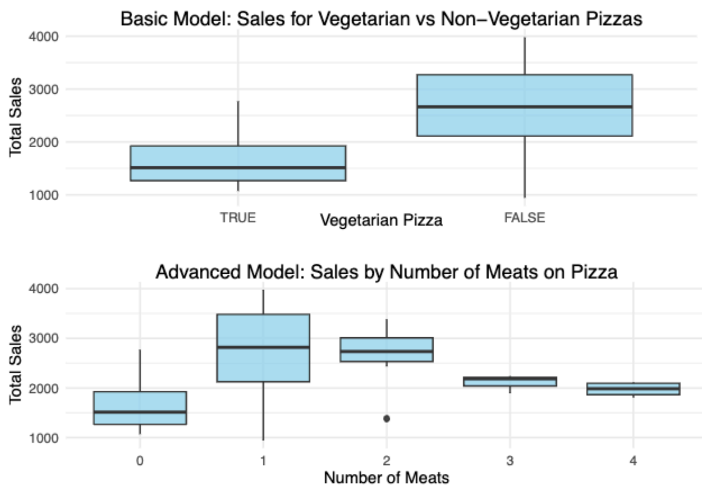

Figure 1: Box plots of aggregated pizza sales for the basic and advanced models

In the bottom half of *Figure 1*, the box plot summarizes Total Sales distribution based on Meat Count. The median Total Sales increases up to 2 meats, reaching a peak, and declines beyond that point. Despite a positive linear trend between Meat Count and Total Sales, the non-monotonic relationship indicates an optimal range for maximizing Total Sales. This observation led to the inclusion of a Meat Count squared variable in our regression model.

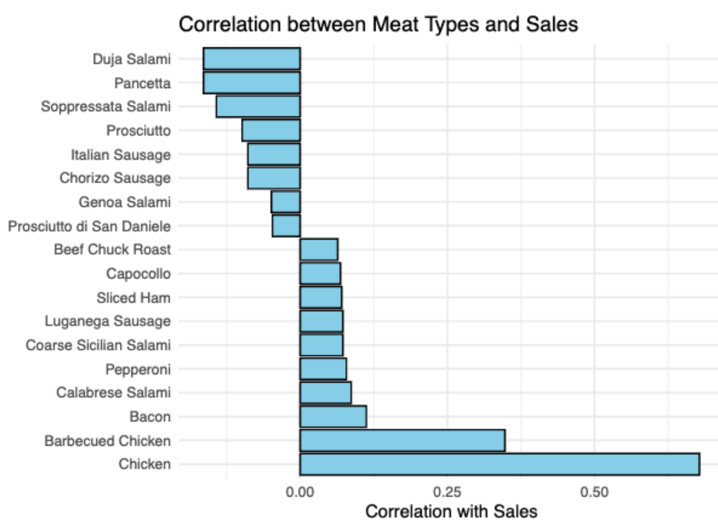

Figure 2: Summary of correlation value between specific meat type and total sales in decending order

Figure 2 shows the calculated correlation of each individual meat on total monthly sales. Here we can see chicken and barbecued chicken show strong positive correlations with total sales whereas, soppressata salami, pancetta, and duja salami exhibit significant negative correlations with sales. This may suggest there is significance in the relationship between type of meat and sales, not just number of meats. From this discovery, we decided to explore the impact of specific types of meats on sales in the advanced model below.

Results

Table 1 provides a comprehensive summary of our basic, intermediate, and advanced regression models. Our basic model included a vegetarian indicator, ingredient count, and month information, while the advanced version included linear and quadratic terms for the number of meats on a pizza, as well as accounting for alternate sauces and alternate cheeses. The advanced model included each individual meat as an indicator variable (18 total meats). Across all models, the addition of meats had a highly statistically significant impact on total sales.

Table 1: Estimated Regressions

	Dependent Variable:		
	Total Sales ($)		
	(1)	(2)	(3)
Vegetarian	-306.510^{***}		
	(100.678)		
Meat Count		401.046^{***}	
		(113.280)	
Meat Count^2		-90.620^{***}	
		(33.070)	
Ingredient Count	12.944	36.592	8.348
	(30.343)	(32.096)	(25.436)
Alternate Sauce		-81.133	-688.228^{***}
		(123.225)	(95.382)
Alternate Cheese		-326.020^{***}	-440.813^{***}
		(95.856)	(81.300)
Month of Year	-21.707	-23.104	-10.847
	(19.771)	(19.153)	(11.619)
Intercept	$2,155.193^{***}$	$1,872.213^{***}$	$2,116.646^{***}$
	(231.328)	(247.762)	(196.790)
Hot-Coded Month	✓	✓	✓
Individual Meats			✓
Observations	282	282	282
R^2	0.068	0.136	0.701
Adjusted R^2	0.022	0.084	0.667
Residual Std. Error	724.686 (df = 268)	701.702 (df = 265)	423.097 (df = 252)
Note:			*p<0.1; **p<0.05; ***p<0.01

In the basic model, the vegetarian indicator coefficient was observed at -307 (±101), implying that for a hypothetical pizza with a fixed ingredient count, offering a non-meat option will lower the total sales by approximately $307 (±$101) per month compared to a pizza with meat.

For the intermediate model, the coeficients on meat count, meat count squared, and the alternative cheese indicator were all significant. The meat count coeficient was found to be $401 (±113), the meat count squared coeficient was $-91 (±33), and including an alternative cheese was found to decrease sales by -$326 (±96). The interaction of the meat count and squared terms suggests adding more meats to a pizza initially increases sales but the incremental benefit of adding meat decreases and eventually turns negative. As an example, the marginal benefit of going from 0 meats on a pizza to 1 meat is a $\$401 \cdot 1 - \$91 \cdot 1^2 = \$301$ increase in monthly sales. However, transitioning from a pizza with 3 meats to 4 meats results in a marginal decrease of $(\$401 \cdot 4 - \$91 \cdot 4^2) - (\$401 \cdot 3 - \$91 \cdot 3^2) = -\$236$ in totals sales. Based on the model, the optimal number of meats on a pizza is 2.

In the advanced model, given that a statistical test was performed on each individual meat, the p-value threshold was decreased to $p < \frac{0.05}{18}$ or $p < 0.003$ for statistical significance. Given this new threshold value, there were 6 meats associated with an increase in total sales (sliced ham, bacon, chicken, beef chuck roast, capocollo, and prosciutto di salami) and 3 meats associated with a decrease in total sales (Italian sausage, prosciutto, and duja salami). Additionally, alternate sauce now shows a highly significant and large negative impact. This suggests that when you control for *which* meat is on the pizza, changing the sauce is associated with a -\$688 ($\pm$95) decrease in total sales. These changes to the model resulted in a drastic increase in predictive power, increasing from an adjusted R-squared value of 0.09 to 0.67.

Discussion of Limitations

Concerns regarding the i.i.d assumption arise due to several factors. Primarily, there is a time series nature of grouping the total sales by month. Pizza sales of one month could influence pizza sales of the next month due to factors like customer retention or word-of-mouth. We attempt to account for this by including control variables in the model. Specifically, we hot-code the individual months to account for month-to-month variation, and add a month number variable to capture the time ordering component.

Secondly, there is the potential of geographic clustering as we do not have location data in this dataset. Geographical groupings could influence pizza sales by reflecting regional or local sales trends. Additionally, there is the possibility of repeat customers in the database, which may not represent a random sampling and could skew the results towards repeat customers' personal preferences. Finally, the dataset does not specify if promotions or discounts occurred during the sampling. Promotions can change the underlying sampling distribution since a heavier weight with will be applied towards whichever pizza is currently discounted.

Regarding structural limitations, the validity of our estimates on the impact of meat pizza sales may be biased by several omitted variables. An example of such a variable is religion. Many religions tend to restrict meat consumption, which will negatively correlate with the amount meat on pizzas. Additionally, religions tend to emphasize healthier diets and could have a negative correlation with total pizza sales. Therefore, we anticipate a positive omitted variable bias due to religion, which would result in a bias away from zero.

Income level is another variable to consider. Affluent consumers may be able to afford the premium or meat-heavy pizzas, which results in a positive correlation with meat consumption. However, wealthier demographics tend to eat healthier foods, and thus overall pizza sales may decline, resulting in a negative correlation. Therefore, we predict a negative omitted variable bias, resulting in a bias towards zero, underestimating the impact of meat on pizza sales.

One final example - which is challenging to pinpoint bias directionality for - is the impact of marketing. Effective marketing can be assumed to increase the sales of the marketed pizza, regardless of the meat content. We assume this will have a positive correlation with total pizza sales but could have a positive or negative correlation with meat consumption depending on whether meat or vegetarian pizzas were marketed. Therefore, this could lead to a positive or negative bias, depending on which marketing strategy that was employed.

Lastly, while our analysis suggests how certain individual ingredients may impact sales, it does not take into account flavor combinations. Tastes and flavor perceptions are highly individualized and transcend quantitative data, making it challenging for analysis alone to encompass the depth of taste that drive food choices.

Conclusion

This study evaluated the impact of meat on total pizza sales. The regression models shows that pizzas with meat have an increased monthly sales of \$307, and that the optimal number of meats on a pizza is 2. Beyond that value, pizza sales tend to diminish. Additionally, including an alternative cheese is associated with an increase of \$326 in monthly sales.

In future research, we would recommend collecting a more comprehensive dataset to investigate impact of meat on sale for broader topics like regional differences or pizza chain-level differences. The intended goal of this work is to help restaurant owners and chef's curate to drive overall sales for their establishments. Adding more depth to the analysis through those additional variables could enhance the impact of the work.