Do Meat Toppings Affect Pizza Sales?

Datasci 203: Lab Report 2

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

Contents

1 Introduction (1)

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.

2 Data and Methodology (2-3)

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 categories of ingredients were then counted. To further account for time trends for the sales, one hot encoded months and a month variable were included in the models. After organizing the data, it was randomly split into 75% for exploration analysis and 25% of modeling building. The final exploration and model building data set had 102 and 283 data points respectively.

3 An Explanation of Key Modeling Decisions (4)

3.1 Observations Removed:

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

3.2 Variable Transformations:

The dataset is aggregated at the 'pizza_name' and 'month' level without additional changes. These alterations to the data involves creating a summary dataset at a higher level of granularity, providing a monthly overview of key metrics for each pizza type.

3.3 Intentional Covariate Exclusions:

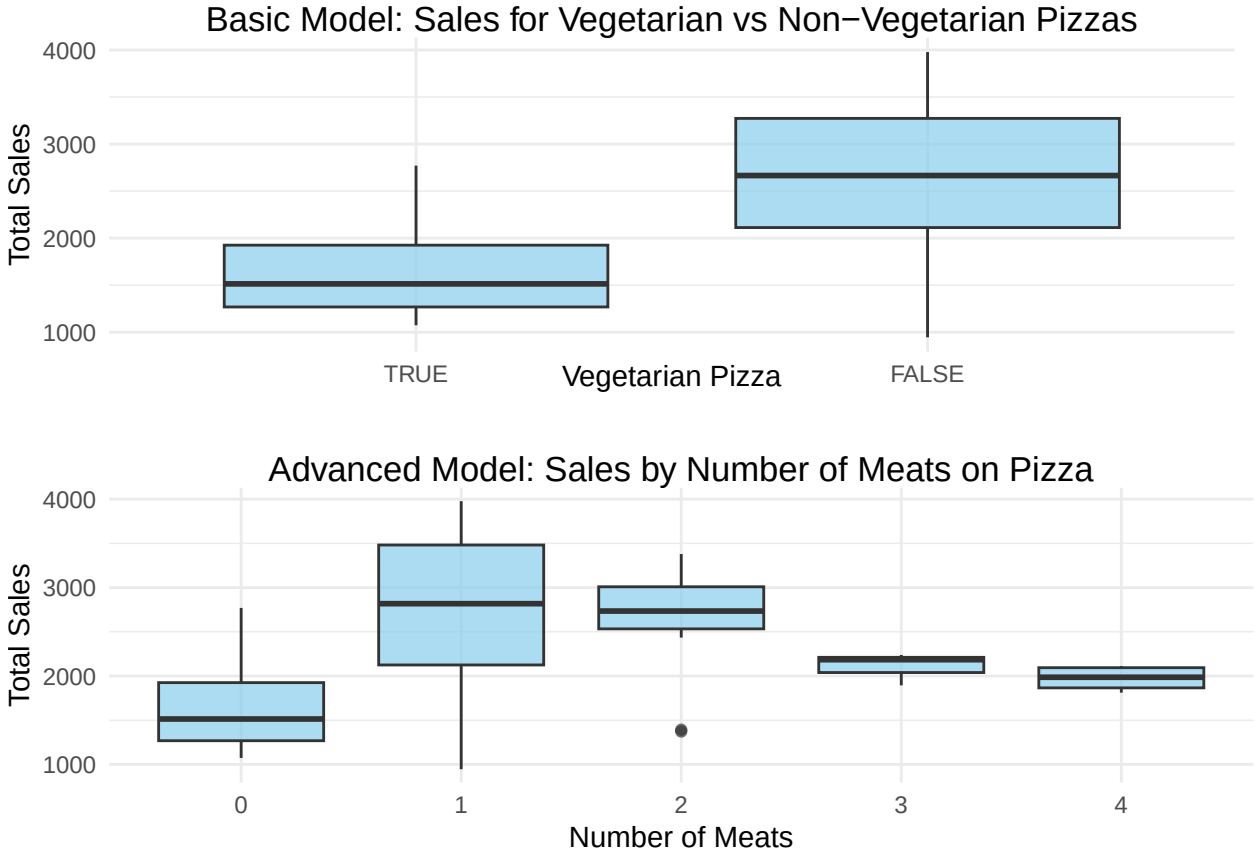
The only covariates that were removed was the month of January as the month were one hot encoded from month_01 to month_12. Otherwise, the transformed dataset includes relevant variables for the analysis, such as counts of ingredients, meat, alternative cheese (not mozzarella), alternative sauce (not red sauce), veggie, and binary indicators. The inclusion of these variables aligns with the research question and allows for a comprehensive analysis of pizza sales based on ingredients in the pizza.

4 A Table or Visualization (5)

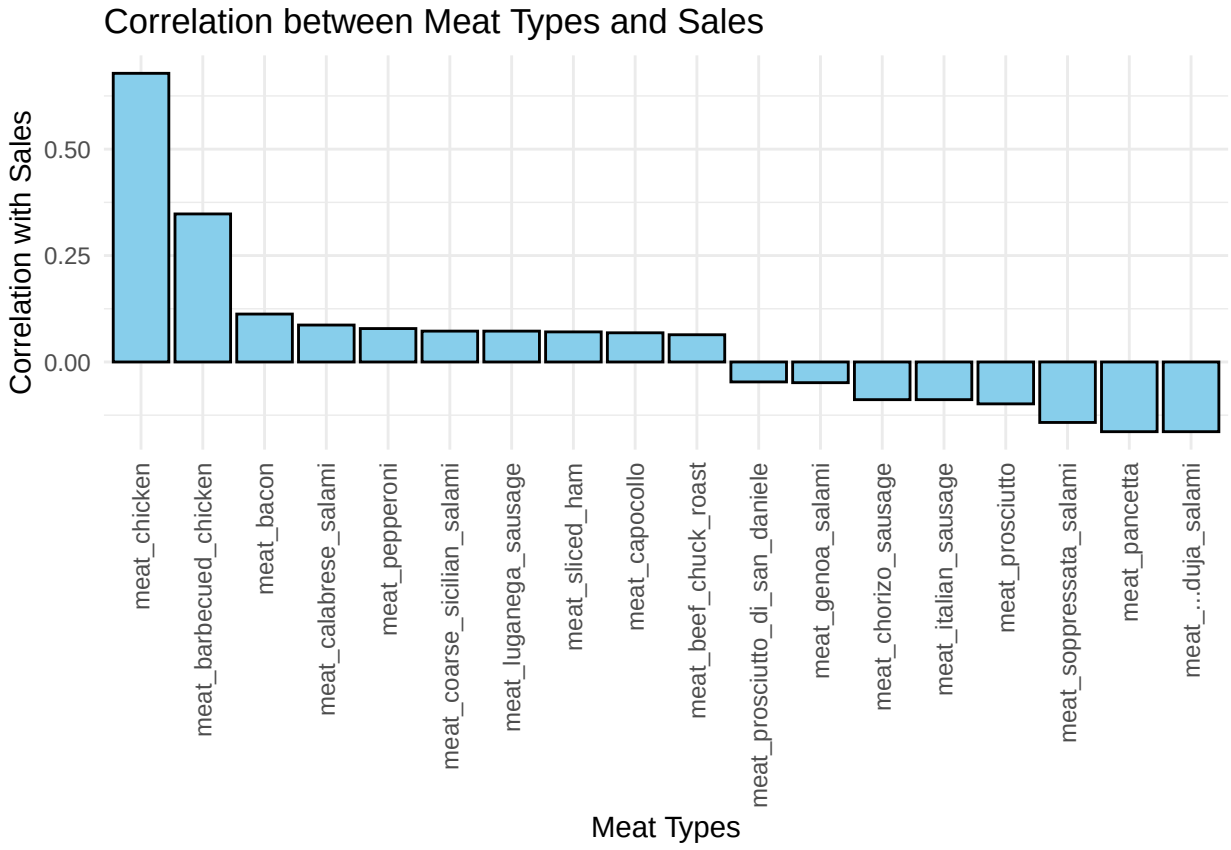
The box plot analysis indicates that, on average, non-vegetarian pizzas tend to have higher Total Sales compared to vegetarian pizzas. While medians provide insights into the central tendency, the spread of the box plot and the presence of outliers suggest that the distribution of Total Sales is broader for non-vegetarian pizzas. This finding could guide further investigations into the factors influencing Total Sales for each category, helping restaurant owners make informed decisions about their pizza offerings.

In the scatterplot with the trendline, we observe a positive linear relationship between Meat Count and Total Sales. The blue regression line indicates a positive slope, suggesting that as the Meat Count increases, the Total Sales tend to increase. However, it's essential to note that the scatterplot points show some variability, and there might be other factors influencing the relationship especially around Meat Count of 2

The boxplot provides a summary of the distribution of Total Sales across different Meat Counts. The boxplot indicates that the median Total Sales increase with the Meat Count up to 2, where it reaches a peak. Beyond 2 meats, the median Total Sales start to decline. This pattern is consistent with the observation in the scatterplot.



The scatterplot and boxplot together indicate that while there is a positive linear trend between Meat Count and Total Sales, the relationship is not strictly monotonic. The plateau and subsequent decline in median Total Sales beyond 2 meats suggest that there may be an optimal range of Meat Count for maximizing Total Sales. Further analysis and potential model refinement may be needed to capture the nuanced relationship between the predictors and Total Sales.



5 Results (6-7)

5.1 Basic Model: Vegetarian or not

5.2 Advanced Model: Number of meats

Table 1 provides a comprehensive summary of our basic 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. Across both models, the addition of meats had a highly statistically significant impact on total sales. In the basic model, the vegetarian indicator coefficient was observed at -307, implying that for a hypothetical pizza with a fixed ingredient count, offering a non-meat option will lower the total sales by approximately \$307 per month compared to a pizza with meat.

For the advanced model, both the linear and quadratic meat count terms were highly significant, as well as the alternate cheese indicator. The linear term for meat count has positive coefficient, while the squared term has a negative coefficient. The interaction of these terms suggests that initially, adding more meats to a pizza increases sales, but after a certain number of meats, the incremental benefit decreases and eventually turns negative. To frame this in an example, the marginal benefit of going from 0 meats on a pizza to 1 meat on a pizza is $\$401 \cdot 1 - \$91 \cdot 1^2 = \$301$ increase in monthly sales. However, transitioning from a pizza with 3 meats to a pizza with 4 meats, results in a marginal difference of $(\$401 \cdot 4 - \$91 \cdot 4^2) - (\$401 \cdot 3 - \$91 \cdot 3^2) = -\$236$ decrease in totals sales. Based on the model, the optimal number of meats on a pizza is 2, as any increase beyond this number is associated with a negative impact on total sales.

Additionally, the alternate cheese indicator coefficient was highly significant with a value of -\$326, indicating that pizzas with cheese other than mozzarella sell \$326 less per month than pizzas with mozzarella cheese. Notably, the month of sale did not have any significant impact on total sales.

Table 1: Estimated Regressions

	Dependent Variable:	
	Total Sales ($)	
	(1)	(2)
Vegetarian	-306.510^{***}	
	(100.678)	
Meat Count		401.046^{***}
		(113.280)
Meat Count^2		-90.620^{***}
		(33.070)
Ingredient Count	12.944	36.592
	(30.343)	(32.096)
Alternate Sauce		-81.133
		(123.225)
Alternate Cheese		-326.020^{***}
		(95.856)
Month of Year	-21.707	-23.104
	(19.771)	(19.153)
Intercept	$2,155.193^{***}$	$1,872.213^{***}$
	(231.328)	(247.762)
Hot-Coded Month	✓	✓
Observations	110	110
R^2	0.157	0.249
Adjusted R^2	0.043	0.119
Residual Std. Error	712.348 (df = 96)	683.213 (df = 93)
Note:		*p<0.1; **p<0.05; ***p<0.01

6 Discussion of Limitations (8)

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.

7 Conclusion (9)

This study evaluated the impact of meat on total pizza sales. The regression models predict that pizzas with meat will increase sales by $307 per month, and that the optimal number of meats on a pizza is 2. Beyond that value, pizza sales will begin to diminish. Additionally, swapping out alternate cheese for mozzarella cheese will reduce sales by approximately $326 per month.

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.