Do Meat Toppings Effect 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. 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 110 and 275 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:

No covariates were intentionally excluded from the provided information. The transformed dataset includes relevant variables for the analysis, such as counts of ingredients, meat, alternative cheese (not mozzarella), alternarive 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.
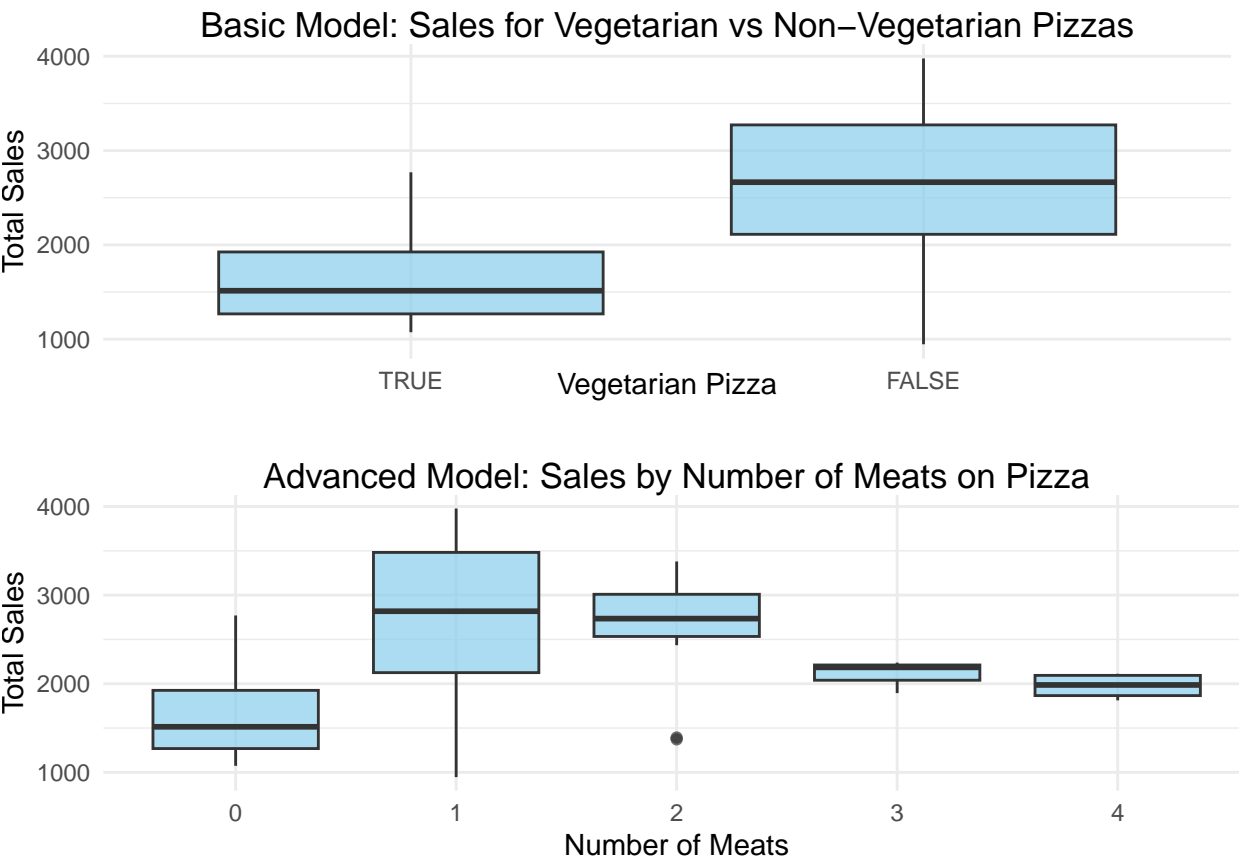
The scatterplot and boxplot together indicate that while there is a positive linear trend between Meat Count and Total Sales, the relationship is not strictly monotonic. The plateau and subsequent decline in median Total Sales beyond 2 meats suggest that there may be an optimal range of Meat Count for maximizing Total Sales. Further analysis and potential model refinement may be needed to capture the nuanced relationship between the predictors and Total Sales.

4.1 Model 3: Sales = Intercept + Indicators_for_each_type_of_meat + Number_of_topings + Month + Alternative_Sauce_Indicator + Alternative_Cheese_Indicator (not just mozzarella)

Correlation between Meat Types and Sales
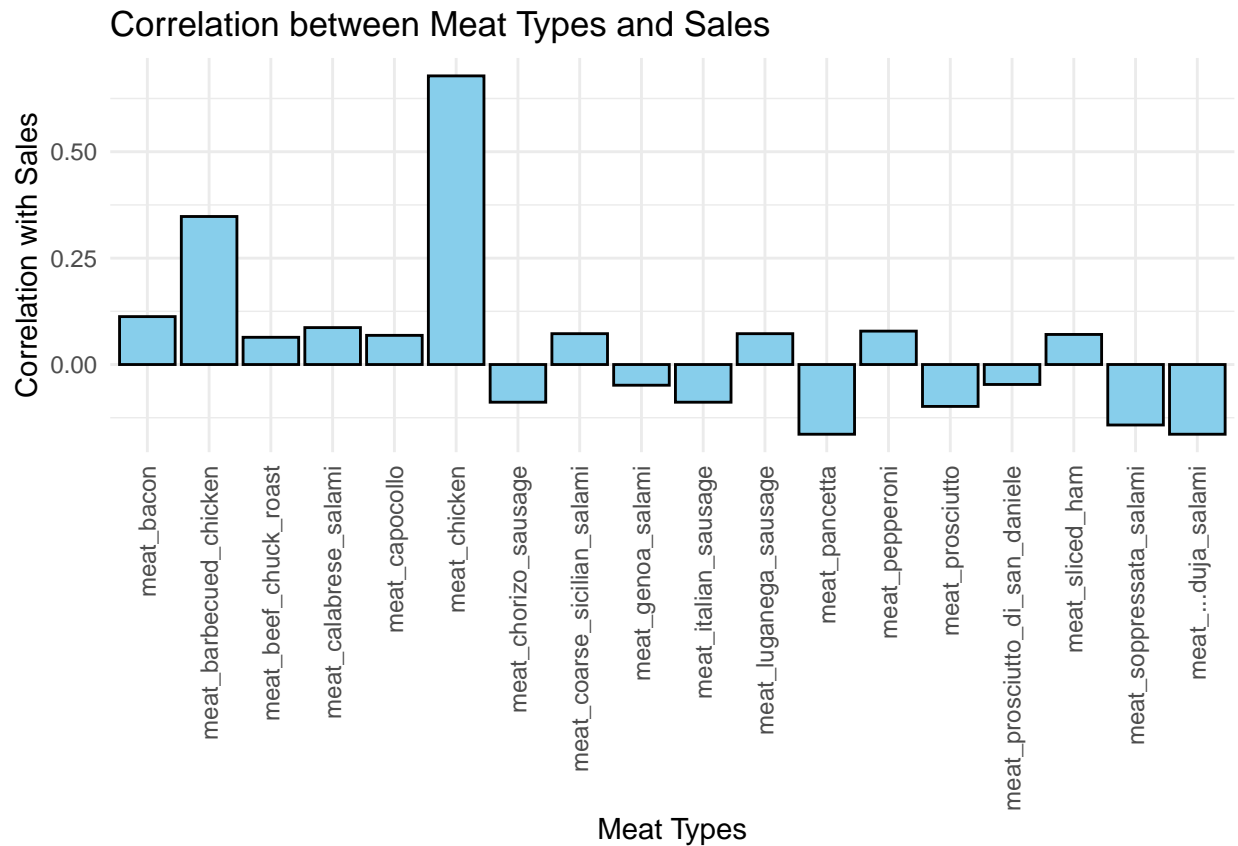
5 Results (6-7)

5.1 Basic Model: Vegetarian or not

```
##
## Call:
## lm(formula = total_sales ~ is_vegetarian + ingredeint_count +
##     month + month_02 + month_03 + month_04 + month_05 + month_06 +
##     month_07 + month_08 + month_09 + month_10 + month_11 + month_12,
##     data = model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1383.40  -589.51    17.13   527.53  1786.55
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2155.193    231.328   9.317  < 2e-16 ***
## is_vegetarianTRUE  -306.510    100.678  -3.044  0.00256 **
## ingredeint_count     12.944     30.343   0.427  0.67002
```

3

```
## month                -21.707     19.771  -1.098  0.27323
## month_02             -149.212    196.483  -0.759  0.44827
## month_03               94.961    189.597   0.501  0.61688
## month_04              -90.224    192.305  -0.469  0.63933
## month_05              220.336    181.631   1.213  0.22616
## month_06                1.016    188.678   0.005  0.99571
## month_07              251.288    177.849   1.413  0.15884
## month_08              114.506    188.009   0.609  0.54301
## month_09             -101.171    193.286  -0.523  0.60111
## month_10              -21.833    205.859  -0.106  0.91562
## month_11              184.972    216.656   0.854  0.39400
## month_12                   NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 724.7 on 268 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.06771,    Adjusted R-squared:  0.02248
## F-statistic: 1.497 on 13 and 268 DF,  p-value: 0.1179
```

5.2 Advanced Model: Number of meats

```
##
## Call:
## lm(formula = total_sales ~ meat_count + meat_count2 + ingredeint_count +
##     alt_sauce + alt_cheese + month + month_02 + month_03 + month_04 +
##     month_05 + month_06 + month_07 + month_08 + month_09 + month_10 +
##     month_11 + month_12, data = model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1454.5  -533.6   -41.8   467.5  1915.2
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1872.213    247.762   7.557 6.75e-13 ***
## meat_count       401.046    113.280   3.540 0.000472 ***
## meat_count2      -90.620     33.070  -2.740 0.006557 **
## ingredeint_count  36.592     32.096   1.140 0.255282
## alt_sauceTRUE    -81.133    123.225  -0.658 0.510843
## alt_cheeseTRUE  -326.020     95.856  -3.401 0.000775 ***
## month            -23.104     19.153  -1.206 0.228782
## month_02        -175.702    190.504  -0.922 0.357210
## month_03         100.731    183.901   0.548 0.584326
## month_04         -68.215    186.644  -0.365 0.715045
## month_05         224.158    176.068   1.273 0.204087
## month_06         -18.818    182.867  -0.103 0.918115
## month_07         240.310    172.480   1.393 0.164707
## month_08          94.036    182.500   0.515 0.606797
## month_09         -88.645    187.305  -0.473 0.636415
## month_10           4.126    199.892   0.021 0.983546
## month_11         211.931    209.878   1.010 0.313520
## month_12              NA         NA      NA       NA
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 701.7 on 265 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1357, Adjusted R-squared:  0.08351
## F-statistic:   2.6 on 16 and 265 DF,  p-value: 0.0008766
```

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 by pizza sales of the next month due to factors like customer retention or word-of-mouth. **Talk about how we try to account for this??**. 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. Repeat pulls may not represent a random sampling and could influence the results by skewing 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.