# 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 dataset 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 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. We performed exploration and model building on a 30% subsample of the data. The remaining 70%, totaling {X rows} was used to generate the statistics in this report.

# 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)

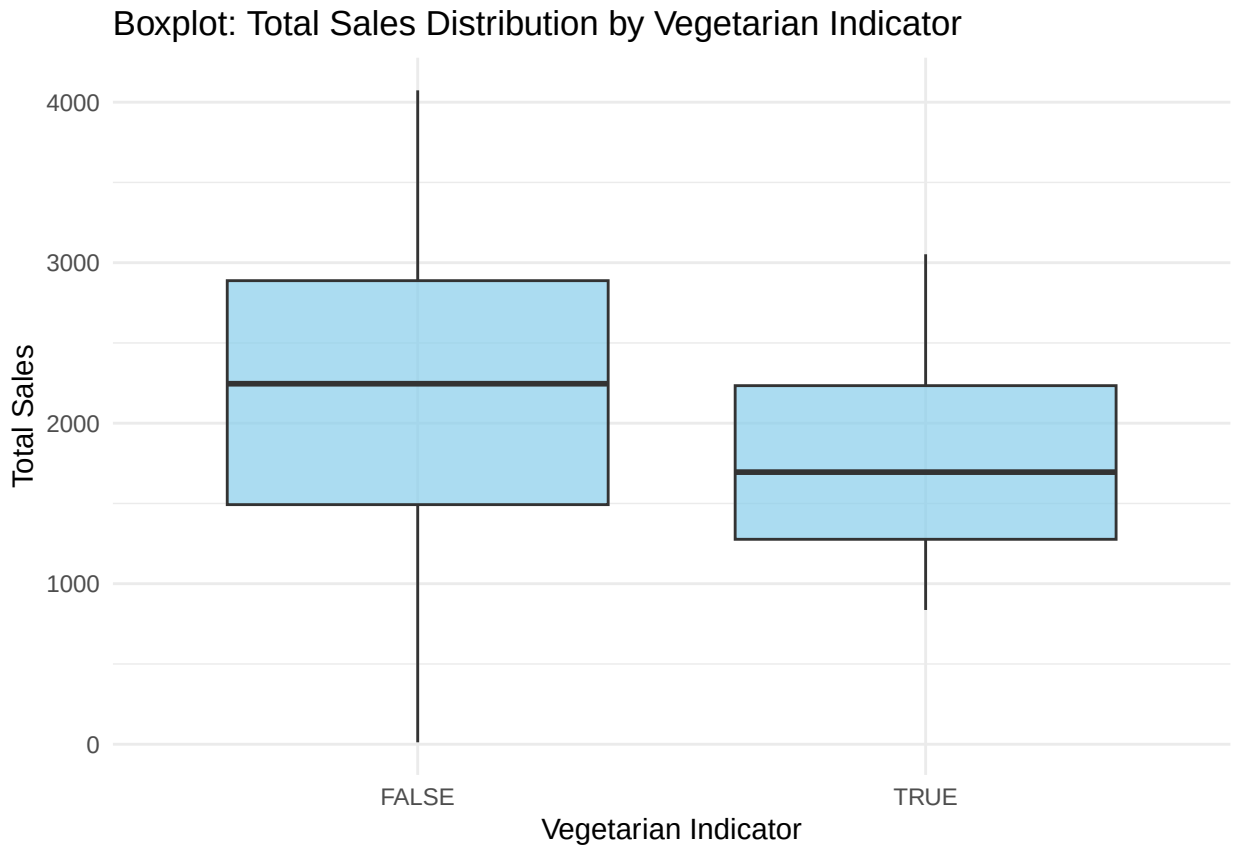
## 4.1 Model 1: Sales = Intercept + Vegitarian_Indicator + Number_of_topings + Month

### 4.1.1 Boxplot: Total Sales Distribution by Vegetarian Indicator

```
train <- read_csv("../data/processed/train_data.csv")
```

```
## New names:
## Rows: 275 Columns: 41
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): pizza_name dbl (37): ...1, month_01, month_02, month_03, month_04,
## month_05, month_06, ... lgl (3): is_vegetarian, alt_cheese, alt_sauce
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
ggplot(train, aes(x = factor(is_vegetarian), y = total_sales)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  labs(title = "Boxplot: Total Sales Distribution by Vegetarian Indicator",
       x = "Vegetarian Indicator",
       y = "Total Sales") +
  theme_minimal()
```




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.

## 4.2 Model 2: Sales = Intercept + Number_of_meats + Number_of_topings + Month + Alternative_Sauce_Indicator + Alternative_Cheese_Indicator (not just mozzarella)

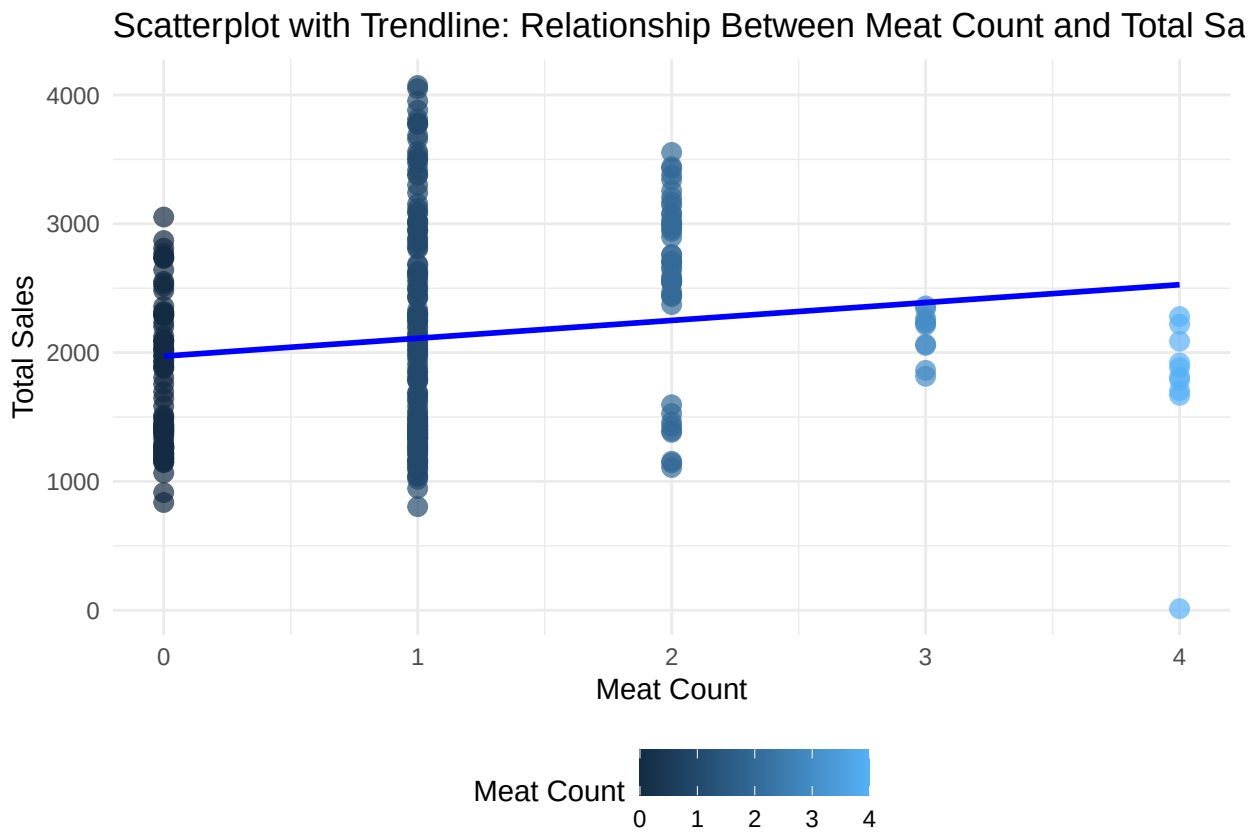
### 4.2.1 Scatterplot with Trendline: Relationship Between Meat Count and Total Sales

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

```
ggplot(train, aes(x = meat_count, y = total_sales, color = meat_count)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, linetype = "solid", color = "blue") +
  labs(title = "Scatterplot with Trendline: Relationship Between Meat Count and Total Sales",
       x = "Meat Count",
       y = "Total Sales",
       color = "Meat Count") +
```

```
  theme_minimal() +
  theme(legend.position = "bottom")
```

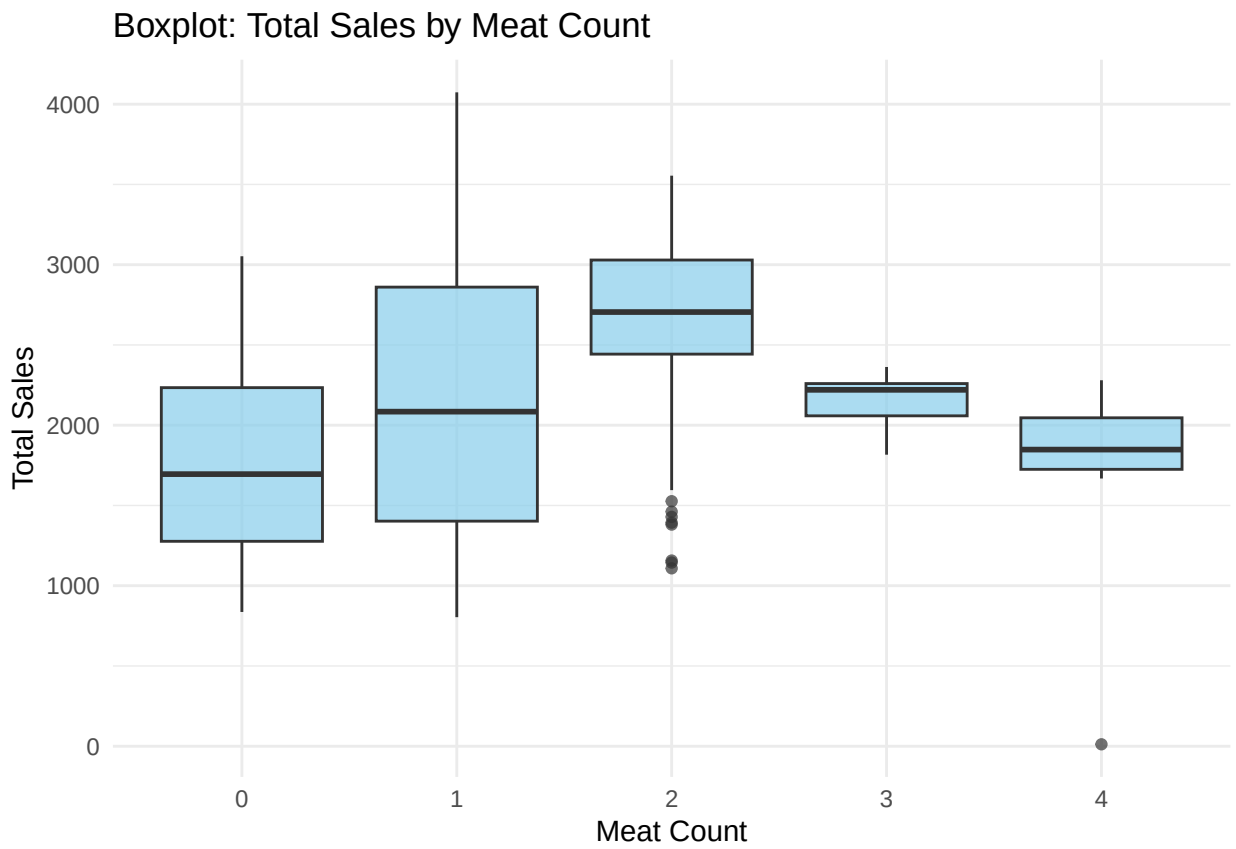
```
## `geom_smooth()` using formula = 'y ~ x'
```




### 4.2.2 Boxplot: Total Sales by Meat Count

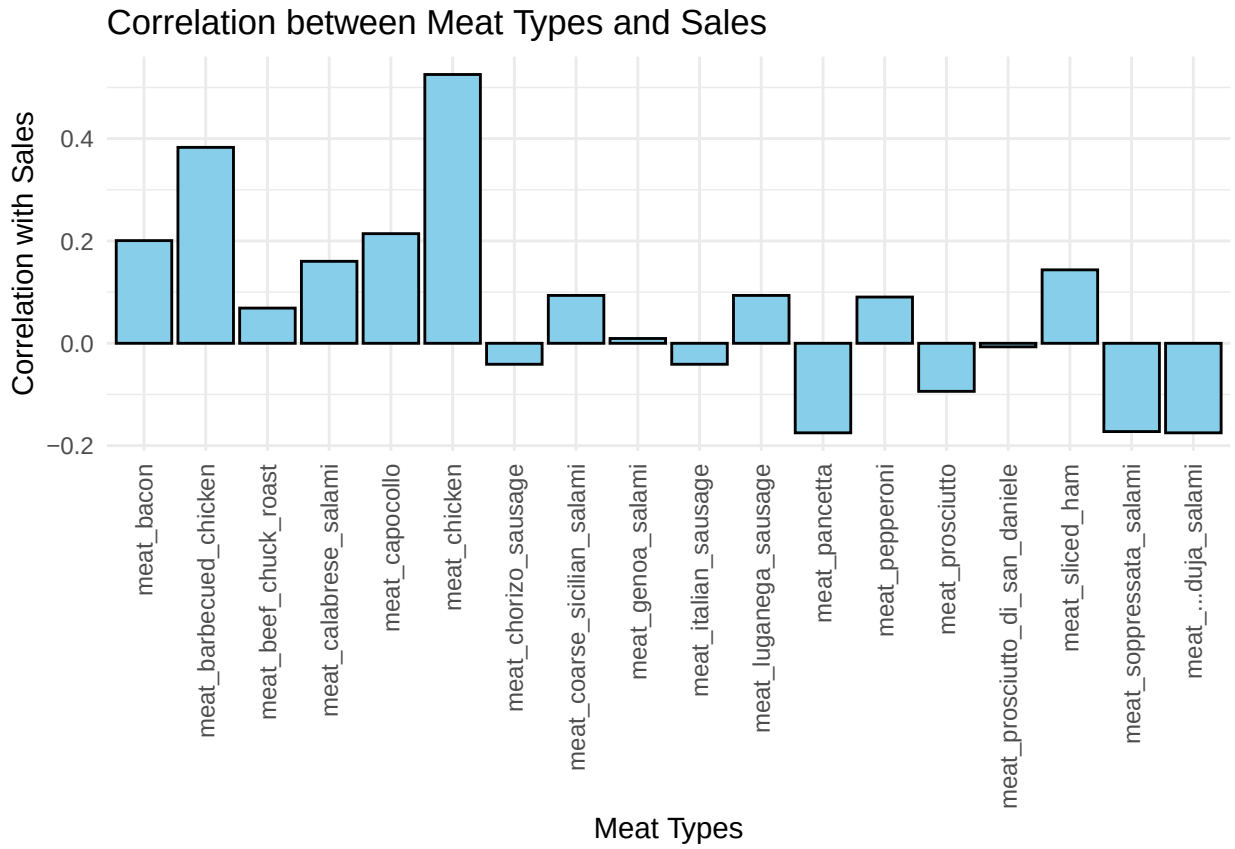
The boxplot provides a summary of the distribution of Total Sales across different Meat Counts. The boxplot indicates that the median Total Sales increase with the Meat Count up to 2, where it reaches a peak. Beyond 2 meats, the median Total Sales start to decline. This pattern is consistent with the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

```
ggplot(train, aes(x = factor(meat_count), y = total_sales)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  labs(title = "Boxplot: Total Sales by Meat Count",
       x = "Meat Count",
       y = "Total Sales") +
  theme_minimal()
```




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.

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




## 5 Results (6-7)

## 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.