# Assignment 1. SQL Analysis Assessment

Yunsik Choung

Master of Science in Business Analytics, Hult International Business School

Business Analysis with Structured Data

Prof. Chase Kusterer

Dec. 9. 2022

## Business Question:

# Do healthier foods cost less?

### 1. Introduce the problem and define "healthy" and "cost".

This study examined those healthier foods cost less by collecting product data from the Whole foods market's website. This study defined *healthier food* as a product with organic certification characteristics, low sodium, low fat, less sugar, gluten-free, and fewer calories to determine the research question. According to McEvoy (2019), USDA-certified organic foods are grown and processed according to federal guidelines addressing, among many factors, soil quality, animal raising practices, pest and weed control, and additives. Moreover, from the Dietary Guidelines for Americans, healthier eating habits come from low-fat, low-sodium, gluten-free, non-alcohol, avoiding added sugar and fewer calories (U.S. Department of Agriculture and U.S. Department of Health and Human Services. 2020). Depending on the above definition, this study focused on the product collected from the website of Whole food. These products have the category of special dietary badges for representing each health information on the detailed site. Consequently, the operational definition of *healthier food* is a product with badges of USDA-certified organic, low-sodium, low-fat, sugar-conscious, gluten-free, and fewer calories.

According to the U.S. Bureau of Labor Statistics, for food average prices, each retail price is converted to an effective price per a standard unit of measure (Bureau of Labor Statistics, U.S. Department of Labor. 2022). For example, the estimated average price per ounce for ground beef is multiplied by 16 and published as an average price per pound. The cost information needs to integrate by a unified measurement level, so this study refers to the Customer Price Index report and unifies the regular price as the exact measurement level, cents per pound.

**2. Answer the business question: Do healthier foods cost less?**

After inspecting descriptive statistics and correlations between price per unit and all other variables, this study built a multiple regression model and checked the differences between each independent variable and dummy groups. The final regression model includes dummy variables of each category (the base dummy is beverage) to determine the category's effect on price per pound. Furthermore, this study includes the total package weight for checking the general common effect on the price. Lastly, to establish the research question, all independent variables for the healthy factors are put into the model.

Final Model:

$$Y = \alpha + \beta_1 X_1 + \sum_{n=1}^{8} \beta_{2n} C_n + \beta_3 X_3 + \sum_{k=1}^{5} \beta_{4k} D_k + \epsilon$$

$Y$: Price per pound (¢)
$X_1$: Unified size (g)
$C_n$: Product Category Group (base is Beverages)
$X_3$: Calories per serving
$D_k$: Dummy variables of health factors

The model is fitted in $p < 0.05$ level, f statistics is 5.302, and $adj\ R^2$ is 0.225, which can explain 22.5% of observations. Organic (63.049), sugar-conscious (168.226), and gluten-free (210.308) coefficients are positive but not significant in $p < 0.05$ level. Low-sodium (-493.529) and Low-fat (-503.062) coefficients are averse to the price per pound and statistically significant in $p < 0.05$. Calories per serving (-1.957) have a negative coefficient on price per pound at a significance of $p < 0.05$.

In conclusion, healthier food, low-fat and low-sodium cases, can cost less. On the other hand, fewer calories per serving of food need more cost; organic and sugar-conscious food and gluten-free food need more cost but not significant statistically.

**3. Provide your top three actionable insights.**

First, Categorize dietary information. Organic USDA-certified, USHHS Healthier foods (include Low-sodium, Low-fat, Less-sugar, Low-Calory, Gluten-free), Special Dietary (Other all Special dietary). The existing presence of dietary information is too distributed at a single level. Some information is good for the customer who has an interest in healthier food, while others are convenient to provide exceptional dietary food for the customer who has a plan for each type of diet. Therefore, categorizing this information based on each purpose give the customer more fitted tips for consuming. This study recommends classifying based on three types, USDA-certified organic, USHHS-suggested healthier food information, and Special Dietary to give customer information more transparent.

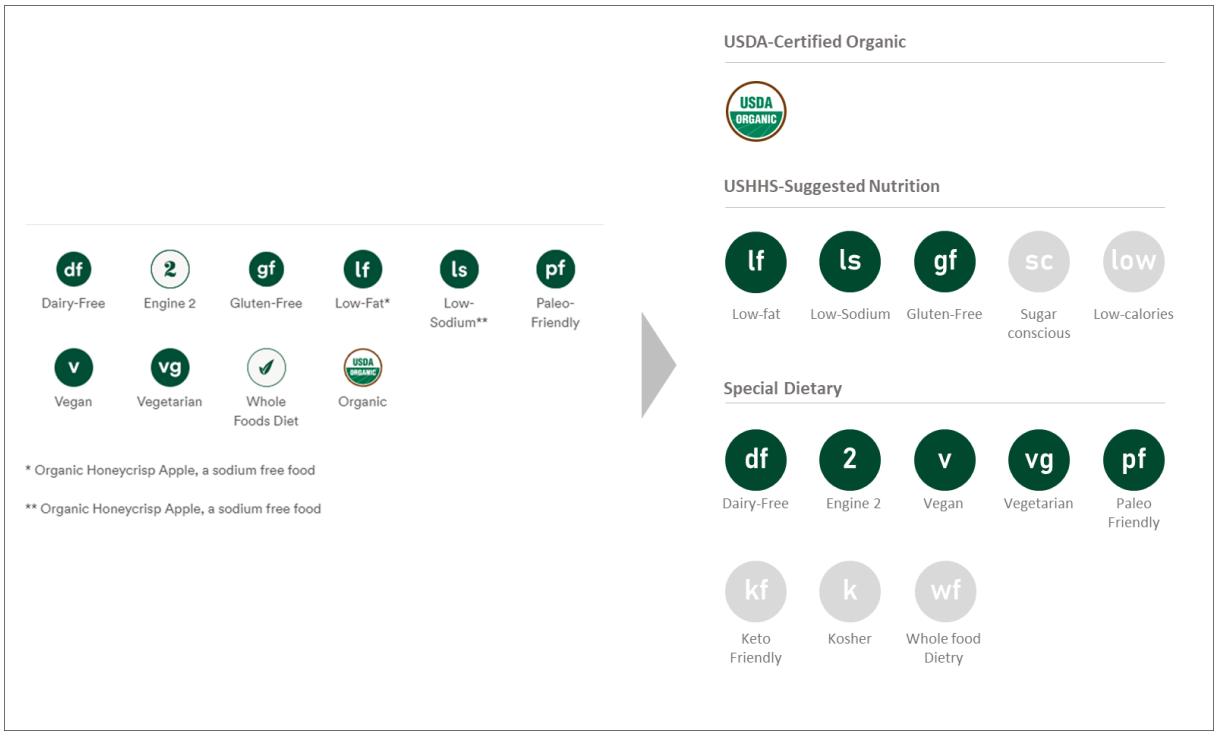


**Figure 1. The new presence of health information recommendations.**

Second, inform the price per same measurement unit and an average price per unit in its product category. According to Whole foods market's website, there are only regular prices and promoted prices on the detailed product page. Consumers cannot compare with other products in the same category with this information. Suppose there is the price per the same measured level and the product's category average price per unit. In that case, consumers could recognize more price information and decide more rationally to choose products.

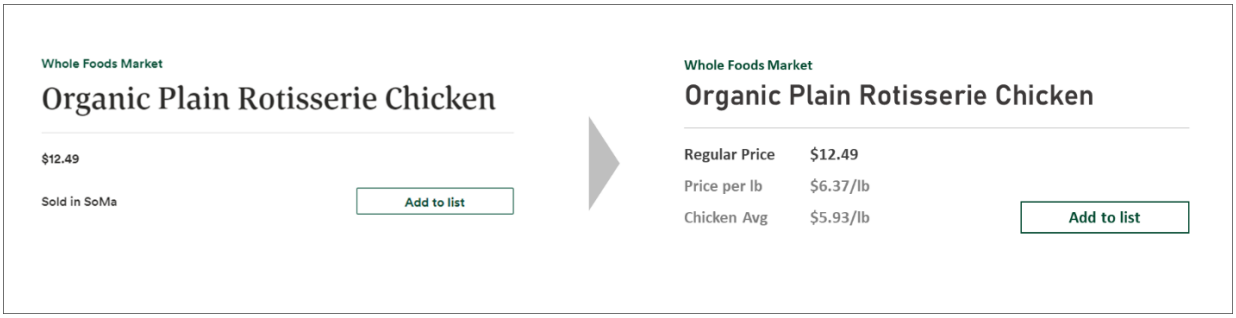


**Figure 2. Giving more price information.**

Third, use strategic healthier factors in customer profiling. This study established the effect of healthier factors on the product price. Some healthy factors have positive efficiency, while others have negative efficiency. If Whole foods market's analysts use this information to analyze consumer behavior, this study expected that this information could provide a new customer segmentation model based on healthier factors. For instance, first, a customer group that has an interest in healthier food but does not matter about the price level; second, a customer group that has an interest in healthier food but is price conscious sensitively; finally, a customer group that does not have any sensitivities in both price and healthy food. For the first group, we can expect to develop a new suggestion algorithm to suggest a similar product with more health factors even though it is more expensive than the current item. For the second group, we can

recommend similar cheaper food with more health factors like low sodium and fat rather than the current product in whose shopping cart. Lastly, we can design a new promotion for not sensitive consumers to inform them of healthier foods and unique dietary plans based on their purchase history.

## References

McEvoy, Miles. (2019). *Organic 101: What the USDA Organic Label Means*, National Organic Program Deputy Administrator in Health and Safety, U.S Department of Agriculture. https://www.usda.gov/media/blog/2012/03/22/organic-101-what-usda-organic-label-means

U.S. Department of Agriculture and U.S. Department of Health and Human Services. (2020). *Dietary Guidelines for Americans, 2020-2025*. 9th Edition. December 2020. https://www.dietaryguidelines.gov/

Bureau of Labor Statistics, U.S. Department of Labor. (2022). *CPI: Average price data*. Consumer Price Index. https://www.bls.gov/cpi/factsheets/average-prices.htm

# Appendix

**Table 1. Descriptive Statistics of numeric variables**

| *Category* | *N* | *(%)* | *Mean of Price per lb* | *Mean of Total weight(g)* | *Mean of Calories per lb* | *Mean of Calories per serving* |
|---|---|---|---|---|---|---|
| Produce | 21 | 9.42 | 1535.66 | 150.288 | 259.686 | 50.095 |
| Dairy and Eggs | 31 | 13.90 | 813.778 | 639.077 | 644.686 | 105.032 |
| Meat | 39 | 17.49 | 1715.235 | 352.695 | 1133.709 | 151.333 |
| Prepared Foods | 14 | 6.28 | 1506.049 | 372.149 | 633.395 | 348 |
| Bread Rolls & Bakery | 21 | 9.42 | 929.583 | 376.325 | 1180.486 | 117.857 |
| Desserts | 24 | 10.76 | 1516.467 | 348.152 | 1247.836 | 205.25 |
| Frozen Foods | 22 | 9.87 | 832.597 | 392.095 | 1096.21 | 281.455 |
| Beverages | 22 | 9.87 | 1700.331 | 735.399 | 120.5 | 46.927 |
| Snacks | 29 | 13.00 | 1891.082 | 193.638 | 2065.479 | 129.828 |
| Total | 223 | 100.00 | 1398.821 | 397.361 | 986.214 | 150.1 |

**Table 2. Desciptive Statistics of Dummy Variables.**

| *Category* | *N* | *(%)* | *Mean of is liquid* | *Mean of Gluten-fee* | *Mean of organic* | *Mean of sugar-conscious* | *Mean of Low-Sodium* | *Mean of Low-fat* |
|---|---|---|---|---|---|---|---|---|
| Produce | 21 | 9.42 | 0 | 1 | 0.619 | 0.571 | 0.952 | 1 |
| Dairy and Eggs | 31 | 13.90 | 0.29 | 0.226 | 0.387 | 0.323 | 0.581 | 0.161 |
| Meat | 39 | 17.49 | 0 | 0.205 | 0.333 | 0.59 | 0.385 | 0.308 |
| Prepared Foods | 14 | 6.28 | 0 | 0.214 | 0.143 | 0.357 | 0.071 | 0 |
| Bread Rolls & Bakery | 21 | 9.42 | 0 | 0.143 | 0.048 | 0.333 | 0.143 | 0.524 |
| Desserts | 24 | 10.76 | 0.5 | 0.292 | 0.167 | 0 | 0.333 | 0.042 |
| Frozen Foods | 22 | 9.87 | 0.5 | 0.227 | 0.273 | 0.182 | 0.364 | 0.136 |
| Beverages | 22 | 9.87 | 1 | 0.182 | 0.591 | 0.136 | 0.591 | 0.636 |
| Snacks | 29 | 13.00 | 0 | 0.586 | 0.483 | 0.448 | 0.517 | 0.207 |
| Total | 223 | 100.00 | 0.242 | 0.336 | 0.35 | 0.345 | 0.453 | 0.327 |

**Table 3. Correlation between price and other factors.**

| *Factor. 1* | *Factor. 2* | *Coefficient* | *df* | *t-statistics* | *Significant Score* |
|---|---|---|---|---|---|
| Price per lb | Total Weight (g) | -0.353 | 221 | -5.601 | p < .05 |
| | Calories per lb | 0.125 | 221 | 1.869 | p < .10 |
| | Calories per serving | -0.144 | 221 | -2.168 | p < .05 |
| | Organic | 0.068 | 221 | 1.017 | |
| | Low-fat | -0.106 | 221 | -1.586 | |
| | Low-sodium | -0.137 | 221 | -2.057 | p < .05 |
| | Sugar-conscious | 0.094 | 221 | 1.397 | |
| | Gluten-free | 0.128 | 221 | 1.92 | p < .10 |

**Table 4. Mean Comparison Analysis and t-test.**

| *Factors* | *Category* | *N* | *Price per lb* | *Difference* | *t-statistics* | *Significant Score* |
|---|---|---|---|---|---|---|
| Calories Group: Over median = 1 | 1 | 118 | 1208.293 | -404.646 | -2.253 | p < .05 |
| | 0 | 105 | 1612.939 | | | |
| Gluten Free: Over median = 1 | 1 | 75 | 1641.399 | 365.506 | 1.92 | p < .10 |
| | 0 | 148 | 1275.893 | | | |
| Low Fat | 1 | 73 | 1193.789 | -304.815 | -1.586 | |
| | 0 | 150 | 1498.604 | | | |
| Low Sodium | 1 | 101 | 1195.77 | -371.151 | -2.057 | p < .05 |
| | 0 | 122 | 1566.921 | | | |
| Organic | 1 | 78 | 1524.26 | 192.916 | 1.017 | |
| | 0 | 145 | 1331.344 | | | |
| Sugar-conscious | 1 | 77 | 1572.504 | 265.283 | 1.397 | |
| | 0 | 146 | 1307.221 | | | |
| Total Weight Group Over median = 1 | 1 | 103 | 722.094 | -1257.58 | -7.81 | p < .05 |
| | 0 | 120 | 1979.678 | | | |

**Table 5. Multiple Regression Models on Price per lb.**

| Variables | Coefficients | Standard Error | t Stat | P-value | Significant Score |
|---|---|---|---|---|---|
| Intercept | 2,943.94 | 341.631 | 8.617 | 0.000 | $p < 0.05$ |
| Total Weight (g) | -0.868 | 0.162 | -5.359 | 0.000 | $p < 0.05$ |
| Produce Dummy | -552.14 | 431.369 | -1.28 | 0.202 | |
| Dairy and eggs Dummy | -1,128.24 | 352.281 | -3.203 | 0.002 | $p < 0.05$ |
| Meat Dummy | -445.089 | 359.443 | -1.238 | 0.217 | |
| Prepared foods Dummy | -512.622 | 511.049 | -1.003 | 0.317 | |
| Bread and bakery Dummy | -1,212.10 | 395.346 | -3.066 | 0.002 | $p < 0.05$ |
| Desserts Dummy | -609.896 | 397.501 | -1.534 | 0.126 | |
| Frozen foods Dummy | -1,067.61 | 423.621 | -2.52 | 0.012 | $p < 0.05$ |
| Snacks and chips Dummy | -500.435 | 378.179 | -1.323 | 0.187 | |
| Calories per serving | -1.957 | 0.877 | -2.231 | 0.027 | $p < 0.05$ |
| Organic Dummy | 63.046 | 184.119 | 0.342 | 0.732 | |
| Low-fat Dummy | -503.062 | 224.525 | -2.241 | 0.026 | $p < 0.05$ |
| Low-sodium Dummy | -493.529 | 185.687 | -2.658 | 0.008 | $p < 0.05$ |
| Sugar-conscious Dummy | 168.226 | 192.193 | 0.875 | 0.382 | |
| Gluten-free Dummy | 210.308 | 204.934 | 1.026 | 0.306 | |

N: 223, adjusted R-squared: 0.225, F-statistics: 5.302 ($p < 0.05$)