To: Mr. Mohamed, Executive Committee Member of Janzen Consulting Group

From: Ryan Essem

12/12/2021

The Relationship between Avocado Bag Size, Volume and Average Price

When it comes to purchasing food products, normally the average price of an item depends on its quantity. We tend to believe that buying products in larger quantities and volumes is associated with saving money in the long run and receiving more with your money. According to Bloomberg, avocados are 25% cheaper when buying wholesale and in 2018, avocados cost 17% less at retail outlets.[1] Of course, buying in bulk doesn't always mean you spend less, as our experiences shopping at Costco can explain. However, there seems to be some sort of correlation between bag sizes, total volume and average prices of avocados which would be of interest to the general population as a whole.

You declared that I should investigate the correlation between the bag sizes, total volume and the average price of avocados in order to offer "*BestFoods International"* the best professional advice and actionable solutions in maximizing their profit. I have conducted this case study, re-analyzed the data and want to disclose the results with you Mr. Mohamed. I hypothesized that there is a relationship between at least one of the independent variables and the average price of avocados. This memo will show that there is a relationship between small bags, large bags, x-large bags, total volume and average price of avocados.

To investigate this connection, I used a dataset with information of historical data on avocado prices and sales volume in multiple U.S. markets from 2015 to 2018. In the dataset there were over 18,200 sales represented, I focused on bag sizes, total volume and average prices of these avocado sales. I standardized the small bags, large bags, x-large bags, and average price variables so that the data would be internally consistent and avoid the risk of producing misleading results and statistical significance.

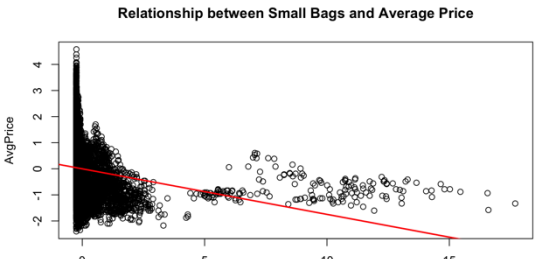
Table 1: Summary Statistics

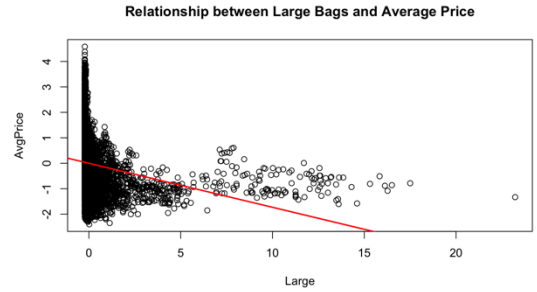
	Freq	Min	Max	Med.	Mean	SD
Small	18249	-0.2442	17.6933	-0.2088	0	1
Large	18249	-0.2227	23.2195	-0.2119	0	1
XLarge	18249	-0.1756	31.0061	-0.1756	0	1
TotalVolume	18249	-0.2463	17.8527	-0.2152	0	1
AvgPrice	18249	-2.39889	4.57941	-0.08935	0	1

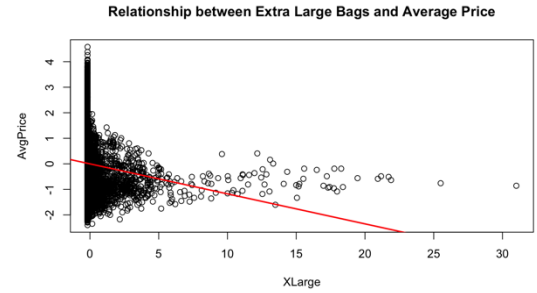
In this case study, I focused on five variables, small bags, large bags, x-large bags, total volume and average price. Small bags, an independent variable, is represented as an integer as the total number of small bag avocados sold. Large bags, an independent

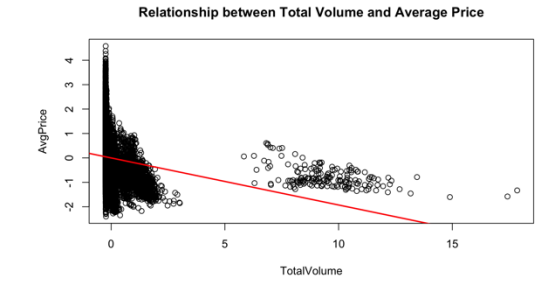
[1] Leslie Patton. (2018). Avocados on Everything? Restaurants Capitalize on Rising Supply. Retrieved from https://www.bloomberg.com/news/articles/2018-05-24/avocados-on-everything-restaurants-capitalize-on-rising-supply

variable, is represented as an integer as the total number of large bag avocados sold. X-large bags, an independent variable, is represented as an integer as the total number of extra-large bag avocados sold. Total volume, an independent variable, is represented as an integer as the total number of avocados sold. Lastly, average price, is represented as an integer as the average price of a single avocado which is still expressed when multiple avocados in a bag are sold. All five variables were collected on a sample size of 18,249 sales.









Regression Model Equation: y = 3.359e^16 + 1.926e^1X – 6.510e^2X + 3.092e^2X – 3.448e^1X + Ei

For the regression analysis, I chose an alpha level of 0.05. The regression models seemed to be fine. The residuals vs fitted plot seemed to show linearity with its flat and straight appearance for the most part until it experienced a peak and then an extreme dip. That distinct change in pattern could possibly mean that there was a slight non-linear relationship between one of the independent variables and the dependent variable, average price. The Q-Q plot seems to be fine as all of the points are over the line, with few deviations in the upper tails. Ultimately, the data looks to be normally distributed and fitted well on the line. The scale-location plot seemed to be pretty linear for the most part, however, it experienced a slight peak followed by a dip and an extreme trend upwards. The residuals vs leverage plot starts off with a non-linear appearance with a curved dip downwards followed by a positive, consistent linear trend. In addition, residuals vs leverage plot seemed to show a lot of high residuals but very few high leverage points meaning there weren't many extreme predictor x values. However, there are three outliers present in the plot.

After analyzing the results, the findings show that there is a highly statistical significance between the independent variables and the dependent variable, average price $R^2 = 0.040$, $F(4, 18244) = 192.2$, $p < 0.05$, power was equal to 1. The r-squared value of 0.040 indicates that the independent variables explain 4% of the variation of the average price of avocados. Therefore, we can reject the null hypothesis and go with the alternative hypothesis which states there is a relationship between the independent variables and the dependent variable, average price. Not only is one of the independent variables statistically significant, but, all four independent variables are statistically significant ($p < 0.05$). The coefficient of -1.9 means that for every one standard deviation increase in the

amount of small bags sold, corresponds to a decrease of about 1.9 U.S. dollars in average price. The coefficient of -6.5 means that for every one standard deviation increase in the amount of large bags sold, corresponds to a decrease of about 6.5 U.S. dollars in average price. The coefficient of 3.1 means that for every one standard deviation increase in the amount of extra-large bags sold, corresponds to an increase of about 3.1 U.S. dollars in average price. Lastly, the coefficient of -3.4 means that for every one standard deviation increase in the amount of the total volume of avocados sold, corresponds to a decrease of 3.4 U.S. dollars in average price. The low R^2 value possibly indicates that the independent variables don't explain much of the variation in the dependent variable, average price, regardless of significance. In other words, while statistically significant, the independent variables small bag, large bag, extra-large bag and total volume are not substantially significant and are not good predictors for the average price of avocados.

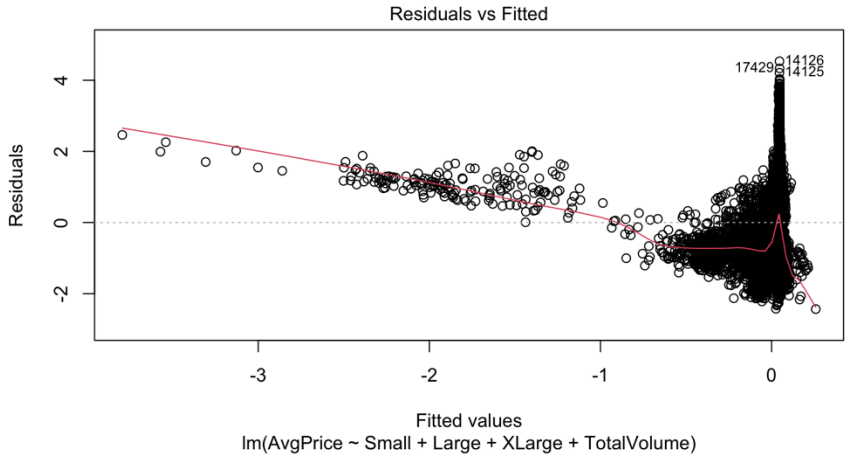
Yet, this study has its limitations. The dataset included dates of avocado sales at national retail volume over a three-year span. However, in the dataset there are sales dates that are out of order and there seems to be omission of dates present. These dates are not sequential because avocado sales were not recorded daily. Due to the lack of daily avocado sales being reported, there can be discrepancies and inaccuracies in the dataset that could've affected this case study's results. We must also consider this dataset stopped collecting sales information with the year 2018. I believe if the avocado sales were recorded daily and diligently with no omissions or miscounts, that an updated dataset could've potentially shown a shift in results.

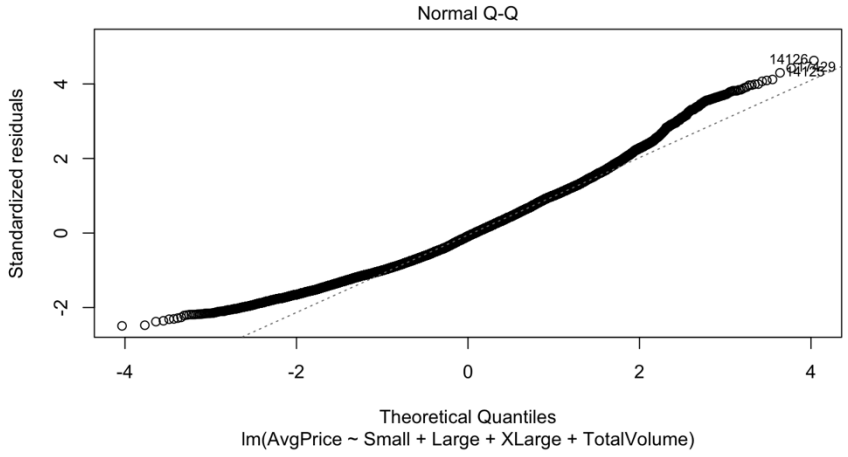
Stargazer Table:

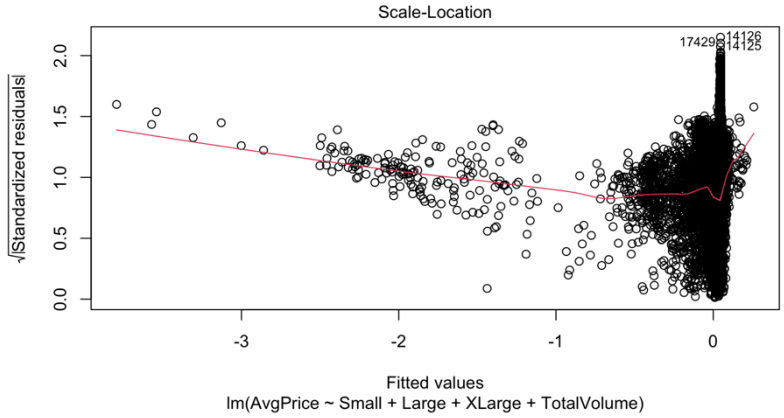
Avocado Bag Size, Bag Volume and Average Price Analysis

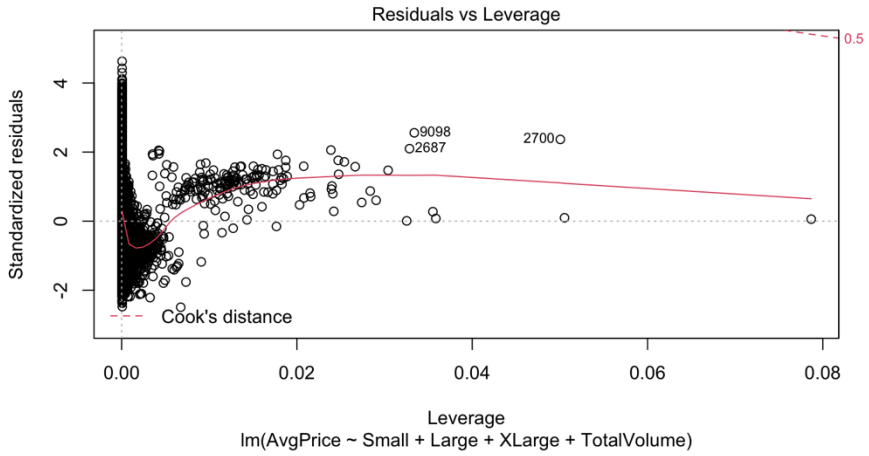
	DV: Average Price of Avocados
	AvgPrice
Small	0.193***
	(0.036)
Large	-0.065***
	(0.017)
XLarge	0.031**
	(0.013)
TotalVolume	-0.345***
	(0.029)
Constant	0.000
	(0.007)
Observations	18,249
R^2	0.040
Adjusted R^2	0.040
Residual Std. Error	0.980 (df = 18244)
F Statistic	192.187*** (df = 4; 18244)
Signifance levels	*p<0.1; **p<0.05; ***p<0.01

Regression Diagnostics:









Correlation Plot: