University of Washington

Avocado Price Elasticity Analysis

Sam Chiang, Eddie Wu, Roy Wang
BUS AN 514
Hema Yoganarasimhan
February 23, 2020

Objectives

There are five types of avocado products: PLU 4046,PLU 4225, PLU 4770, Organic, and Hass Bags. PLU 4046 is a 3~5 oz small/medium Hass Avocado. PLU 4225 is a 8~10 oz large Hass Avocado. PLU 4770 is a 10~15 oz extra large Hass Avocado. Bags vary in size, but represent avocados sold in bags containing multiple avocados. There are some business problems we dressed through this project: What are consumer preferences to Avocados by location and product type in the United States? What are the price elasticity of each category as well as the cross price elasticities? How to make market strategy on each category based on the demand and price analysis?

Dataset

The main avocado data is pulled from https://hassavocadoboard.com/, which has been collected since the beginning of 2016 and is continually expanding today. On their website, they say that they are an organization "that equips the entire global industry for success by collecting, focusing and distributing investments to maintain and expand demand for avocados in the United States'. By analyzing the data, we could find a great competitive strategy to distribute investment and expand demand for avocados.

There are 5293 rows in data for 2018. The main columns include:

- Date - the date of the observation
- Year - the year of the observation
- AveragePrice - the average price of a single avocado
- Region - the city or region of the observation

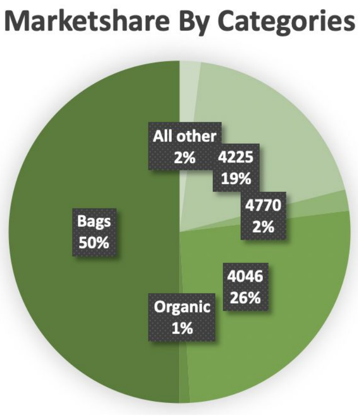
- Total Volume - Total number of avocados sold
- 4046 - Total number of avocados with PLU 4046 sold
- 4225 - Total number of avocados with PLU 4225 sold
- 4770 - Total number of avocados with PLU 4770 sold

Another dataset was mined from the website for elasticity analysis, with 45 rows of data for each region and columns that include:

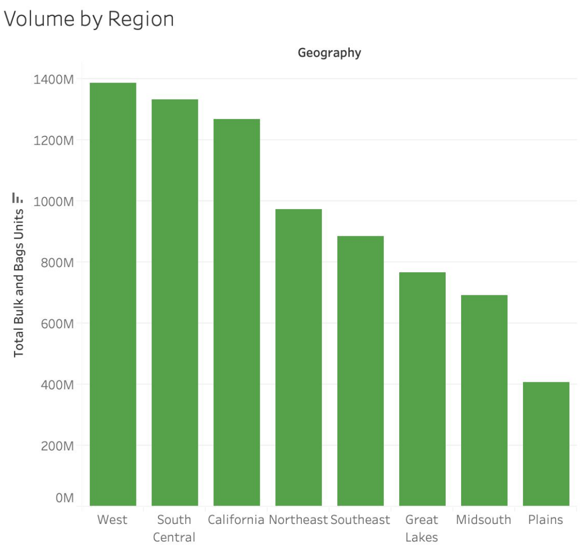
- Year - the year of the observation
- Quarter - the quarter of the observation
- City - the city of the observation
- Quarter - the selling quarter of the observation
- P4046; P4225; P4770; PBags - the prices of the four products
- V4046; V4225; V4770; Vbags - the volume of sales of the products

Exploratory Analysis

The market share chart shows that Organic type of avocado and the 4770 only takes 1% and 2% market share in 2019, respectively. Thus, we decided to exclude these two types of avocados for our price analysis.



We also wanted to sort the regions by its sales volume to determine which regions we want to target. Finding the regions with the highest sales will give us more useful insights about the top avocado consumers. The top four regions in order are West, South Central, California, and Northeast. These four regions will be the basis for our analysis later.



Annual Growth and Demand Volume Visualization

This visualization comes from dataset 1, and we like to understand the sales volume and annual growth in major cities. For the sales volume, we aggregate the weekly sales week together for 2019. For the annual growth, since it's not given, we joined the sales data of 2016 and sales data of 2019 to understand the average 3-year growth in major cities. The code for computing annual growth is below. Once we have all this data, we put them into tableau to generate the visualization.

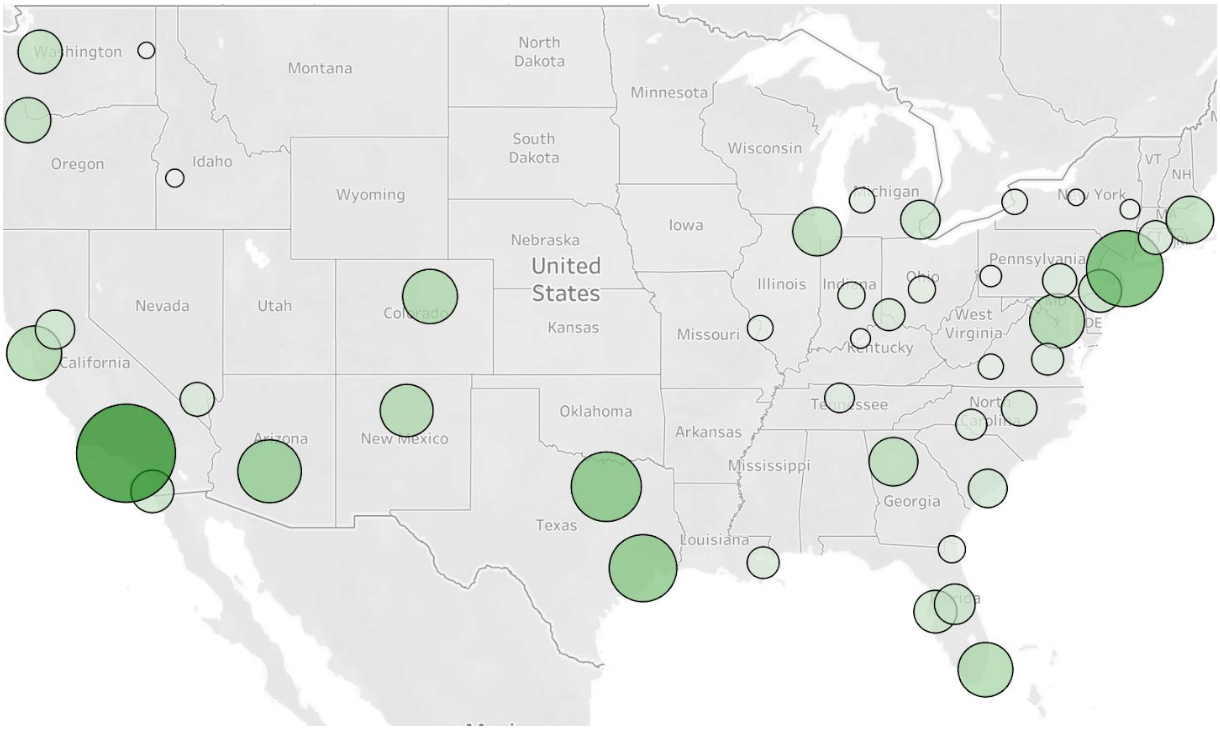
````
```{r}
# cleaning the data set for 2016
Avocado2016<-read.csv("~/Downloads/Avocado_2016.csv")
Avocado2016
Avocado2016sum<- Avocado2016%>%
  group_by(Geography) %>%
  summarize(SumVolume=sum(Total.Bulk.and.Bags.Units))
Avocado2016sum$SumVolume
```
````

````
```{r}
# cleaning the dataset for 2019
Avocado2019<-read.csv("~/Downloads/Avocado2019.csv")
Avocado2019
Avocado2019sum<- Avocado2019%>%
  group_by(Geography) %>%
  summarize(SumVolume=sum(Total.Bulk.and.Bags.Units))
Avocado2019sum
```
````

````
```{r}
#running annual growth
CAGR<-(((Avocado2019sum$SumVolume*1.25)/Avocado2016sum$SumVolume)^(1/3))-1
CAGR
```
````

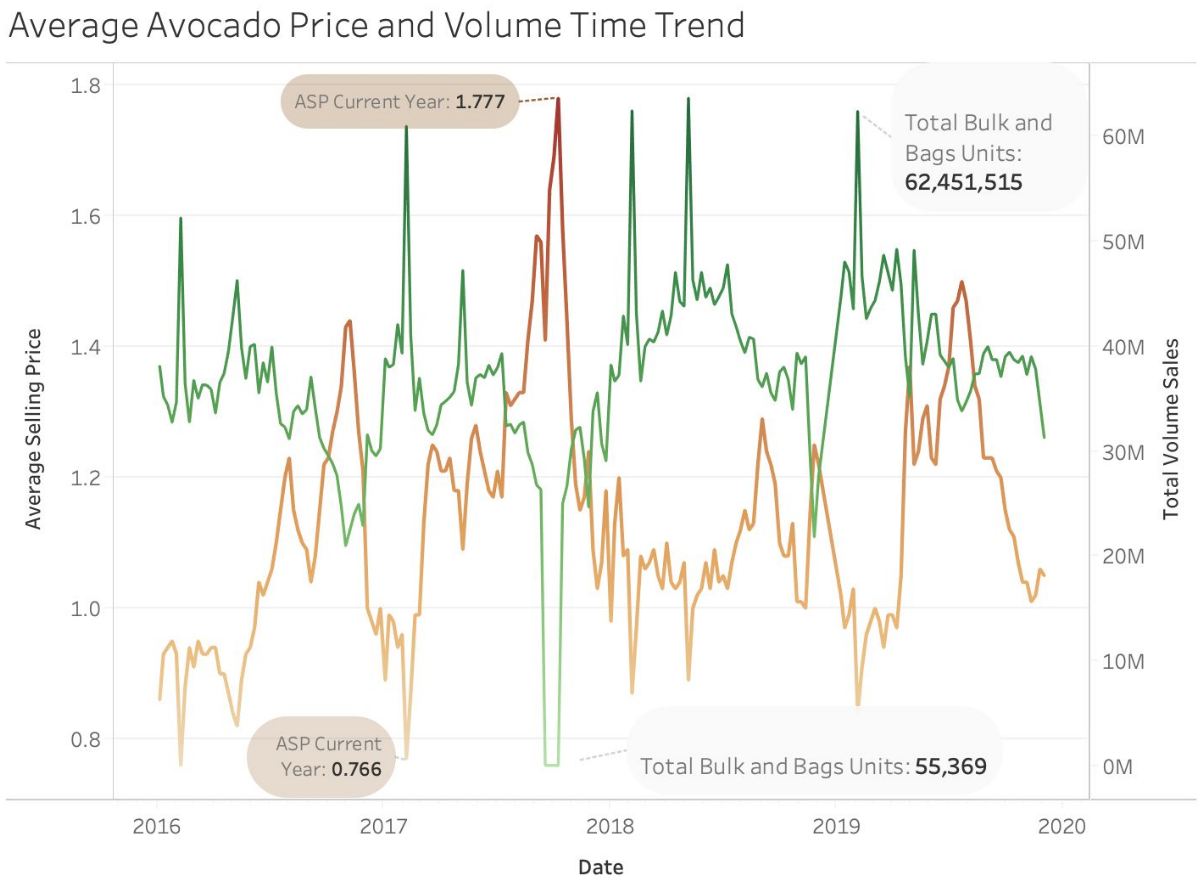
```
 [1] 0.129824786 0.168982944 0.127070598 0.109014417 0.127103901 0.161867982 0.071035408
 [8] 0.208110733 0.054581942 0.130680198 0.172730958 0.145013657 0.079606805 0.185032691
[15] 0.093266058 0.125457600 0.180092147 0.120787328 0.148668478 0.179547178 0.139280708
[22] 0.101464699 0.052715582 0.147955768 0.224431144 0.160559948 0.177027123 0.119028311
[29] 0.169813327 0.151345340 0.164974163 0.188081233 0.165327966 0.098383680 0.181473093
[36] 0.115686844 0.028106430 0.187669422 0.156085440 0.176756320 0.092730695 0.055788208
[43] 0.113070404 0.005484663 0.158969982 0.129650102 0.178097995 0.066840083 0.063781315
[50] 0.158238139 0.183129390 0.120361365 0.080162544 0.075924586
```

Volume Demand of 2019 by City



This visualization is straightforward. This size of the bubble represents the sales volume, where the bigger the bubble, the higher the sales. The color of the bubble represents the annual growth, where the darker the green, the faster the growth.

For the sales volume of 2019, Los Angeles has the highest sales among major cities, accounting for 152.3 M units. While Albany, NY has the lowest sales with 4.3 M units. For the annual growth, every single city has positive growth from 2016 to 2019, ranging from 0.5% to 22%. Cities that are located on the east coast have faster growth than those on the west coast, but the base volume, the size of the circle, of cities on the west coast is so much higher than those on the east coast.



The visualization above shows the average avocado price and volume from 2016 to 2020. Avocado prices see their peak at the beginning of the first quarter of the year, and dips at the end of the last quarter of the year. The highest price of avocados was up to $1.77 and the lowest price seen was $0.766, all within 2017. The highest volume was 62,451,515 in 2019, while the lowest volume was 55,369 in 2017. There was a giant dip in volume with a surge in price in late 2017 due to a weak harvest and insatiable demand for avocados. We saw that customers react highly to an increase in demand: a peak in price shows a mirrored dip in volume. Seeing these price fluctuations, we naturally wanted to analyze the elasticities of avocados.

In-depth Analysis

We applied price elasticity and demand models which we learned in class to analyze the price elasticities and relevant factors impacting demands of avocados. The analytics tools we used in this project are R and Tableau.

To derive the price elasticity of Avocado, we use the following logistic regression to estimate the volume demanded with the product prices as independent variables while also factoring city, year, quarter:

$$\log(Q_i) = +\beta_{i1}\log(P_1) + \beta_{i2}\log(P_2) + \ldots + \beta_{iK}\log(P_K)$$

The Price elasticity (and the corresponding demand) can be categorized as:

1. Inelastic demand: $-1 \leq$ price elasticity < 0
2. Elastic demand: price elasticity < -1

We also use the following formula to calculate the percent change in volume using cross-price elasticities of demand given different pricing scenarios:

$$\%\Delta Q_i = \frac{Q_{i1} - Q_{i0}}{Q_{i0}} \quad = \quad (1+\gamma_1)^{\beta_{i1}}(1+\gamma_2)^{\beta_{i2}}\ldots(1+\gamma_i)^{\beta_{ii}}\ldots(1+\gamma_K)^{\beta_{iK}} - 1$$

Price Elasticity Analysis

For the price elasticity analysis, we used dataset 2 to perform. As mentioned earlier, we decided to focus on products 4225, 4046, and Bags because only these three types avocado have more than 20% the market share, while 4770 and Organic have only 2% and 1% respectively.

We would like to examine how these three types of avocado's price elasticity act differently in different regions. Thus, we picked 4 regions with the highest volume sales among a total of 8 regions, and there are West, Southcentral, California, and Northeast.

First of all, we want to understand own-price elasticity for these three types, and here is a summarized table and screenshots of our code.

````
```{r}
#Bags in California
CAML3<-lm(log(Vbags)~log(P4046)+log(P4225)+log(PBags)+factor(City)+factor(Year)+factor(Quarter),data=
CA)
summary(CAML3)
```

Call:
lm(formula = log(Vbags) ~ log(P4046) + log(P4225) + log(PBags) +
    factor(City) + factor(Year) + factor(Quarter), data = CA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18982 -0.08191 -0.01541  0.08117  0.26247

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               16.12123    0.07795 206.826  < 2e-16 ***
log(P4046)                -0.01796    0.10820  -0.166  0.86915
log(P4225)                -0.06126    0.13832  -0.443  0.66072
log(PBags)                -0.30967    0.09866  -3.139  0.00356 **
factor(City)San Diego     -1.92621    0.05023 -38.351  < 2e-16 ***
factor(City)San Francisco -2.19418    0.06774 -32.389  < 2e-16 ***
factor(Year)2017           0.08794    0.06113   1.439  0.15968
factor(Year)2018           0.53516    0.05167  10.357 6.68e-12 ***
factor(Year)2019           0.82521    0.06374  12.946 1.75e-14 ***
factor(Quarter)2           0.06241    0.05690   1.097  0.28059
factor(Quarter)3           0.07743    0.07439   1.041  0.30551
factor(Quarter)4          -0.05281    0.06577  -0.803  0.42773
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1221 on 33 degrees of freedom
Multiple R-squared:  0.9903,    Adjusted R-squared:  0.987
F-statistic: 304.7 on 11 and 33 DF,  p-value: < 2.2e-16
````

````
```{r}
# 4225 in NorthEast
NorthEastlML3<-lm(log(V4225)~log(P4046)+log(P4225)+log(PBags)+factor(City)+factor(Year)+factor(Quarte
r),data=NorthEast)
summary(NorthEastlML3)
```

Call:
lm(formula = log(V4225) ~ log(P4046) + log(P4225) + log(PBags) +
    factor(City) + factor(Year) + factor(Quarter), data = NorthEast)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49871 -0.14084 -0.04430  0.08971  1.86505

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              16.09747    0.23237  69.276  < 2e-16 ***
log(P4046)                0.34775    0.54321   0.640  0.52647
log(P4225)               -1.26464    0.53930  -2.345  0.02519 *
log(PBags)                0.10471    0.28034   0.374  0.71115
factor(City)New York      0.77757    0.14570   5.337 6.83e-06 ***
factor(City)Philadelphia -0.41418    0.14850  -2.789  0.00871 **
factor(Year)2017         -0.13386    0.18855  -0.710  0.48274
factor(Year)2018          0.02552    0.16291   0.157  0.87648
factor(Year)2019         -0.03889    0.18528  -0.210  0.83502
factor(Quarter)2          0.20387    0.16176   1.260  0.21640
factor(Quarter)3         -0.09638    0.17151  -0.562  0.57796
factor(Quarter)4         -0.30029    0.20223  -1.485  0.14708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3753 on 33 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.6579
F-statistic: 8.693 on 11 and 33 DF,  p-value: 5.868e-07
````

````
```{r}
#4046 in SouthCentral
SouthCentralML3<-lm(log(V4046)~log(P4046)+log(P4225)+log(PBags)+factor(City)+factor(Year)+factor(Quar
ter),data=SouthCentral)
summary(SouthCentralML3)
```

Call:
lm(formula = log(V4046) ~ log(P4046) + log(P4225) + log(PBags) +
    factor(City) + factor(Year) + factor(Quarter), data = SouthCentral)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31709 -0.06196 -0.00317  0.07662  0.22559

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             15.935322   0.155669 102.367  < 2e-16 ***
log(P4046)              -0.104009   0.162325  -0.641   0.5286
log(P4225)               0.204368   0.281238   0.727   0.4754
log(PBags)               0.033221   0.171130   0.194   0.8479
factor(City)New Orleans -1.275848   0.143217  -8.908 1.41e-08 ***
factor(Year)2017        -0.001966   0.086560  -0.023   0.9821
factor(Year)2018         0.113199   0.072615   1.559   0.1340
factor(Year)2019        -0.069284   0.081456  -0.851   0.4046
factor(Quarter)2        -0.023065   0.070691  -0.326   0.7474
factor(Quarter)3        -0.226262   0.086789  -2.607   0.0165 *
factor(Quarter)4        -0.378456   0.072049  -5.253 3.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1325 on 21 degrees of freedom
Multiple R-squared:  0.9773,    Adjusted R-squared:  0.9665
F-statistic: 90.55 on 10 and 21 DF,  p-value: 5.812e-15
````

````
```{r}
#4046 in California
CAML3<-lm(log(V4046)~log(P4046)+log(P4225)+log(PBags)+factor(City)+factor(Year)+factor(Quarter),data=
CA)
summary(CAML3)
```

Call:
lm(formula = log(V4046) ~ log(P4046) + log(P4225) + log(PBags) +
    factor(City) + factor(Year) + factor(Quarter), data = CA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22567 -0.08278 -0.01158  0.07106  0.26348

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               16.49996    0.09209 179.164  < 2e-16 ***
log(P4046)                -0.41993    0.12784  -3.285  0.00242 **
log(P4225)                -0.23869    0.16342  -1.461  0.15359
log(PBags)                 0.22727    0.11657   1.950  0.05976 .
factor(City)San Diego     -2.05552    0.05934 -34.638  < 2e-16 ***
factor(City)San Francisco -1.73234    0.08004 -21.643  < 2e-16 ***
factor(Year)2017           0.14705    0.07222   2.036  0.04984 *
factor(Year)2018           0.06649    0.06105   1.089  0.28398
factor(Year)2019          -0.06669    0.07531  -0.886  0.38229
factor(Quarter)2           0.10716    0.06722   1.594  0.12044
factor(Quarter)3           0.02577    0.08789   0.293  0.77122
factor(Quarter)4          -0.18888    0.07771  -2.430  0.02068 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1442 on 33 degrees of freedom
Multiple R-squared:  0.9801,    Adjusted R-squared:  0.9735
F-statistic: 147.9 on 11 and 33 DF,  p-value: < 2.2e-16
````

Own Price Elasticity in Avocados

Region	Bags (own-price)	4046 (own-price)	4225 (own-price)
Southcentral	-0.42* (mildly inelastic)	-0.1 (very inelastic)	-0.96* (unit elastic)
Northeast	-0.17* (very inelastic)	-0.72 (mildly inelastic)	-1.26* (elastic)
West	-0.2* (very inelastic)	-0.02 (per-inelastic)	-0.56* (mildly inelastic)
California	-0.31 (very inelastic)	-0.42 (mildly inelastic)	-0.51(mildly inelastic)

*Values denoted with * are significant*

Demand Elasticity Model on Bags

Region	Bags(own-price)	4046 (cross-price)	4225 (cross-price)
Southcentral	-0.42* (mildly inelastic)	0.18 (weak impact)	-0.49 (inverse impact)
Northeast	-0.18* (very inelastic)	-0.01 (no impact)	0.39 (positive impact)
West	-0.20 (very inelastic)	-0.24 (inverse impact)	-0.01 (no impact)
California	-0.31* (mildly inelastic)	-0.02 (no impact)	-0.06 (no impact)

*Values denoted with * are significant*

In its own price elasticity in avocados, most of the numbers are significant, meaning this table is valuable.

In these 4 regions, Bags has relatively very inelastic demand, ranging from 0.17 to 0.42. 4046 is more volatile in terms of price elasticity, and only two out of four are showing significance. 4225 is the most elastic product among them.

We also want to examine the impact on the sales of Bags when the price of 4225 and 4046 changes. As a result, the price fluctuations of 4046 and 4225 does not have a huge impact on the sales of the Bags according to the table. From the results, we can conclude that Bags is the flagship category of Avocado.

Price Change Analysis

We chose to separate Bags by itself because it has the largest market share and had the most significant coefficient in our model. We conducted a price change analysis with 4 scenarios:

- All prices increase by 5%
- All prices decrease by 5%
- Bag prices increase by 5%, all else decrease by 5%
- Bag prices decrease by 5%, all else increase by 5%

	BagsPriceChange	4046PriceChange	4225PriceChange	BagsVolumeChange	4046VolumeChange	4225VolumeChange
California	0.05	0.05	0.05	-0.019	-0.021	-0.031
California	-0.05	-0.05	-0.05	0.020	0.022	0.033
California	0.05	-0.05	-0.05	-0.011	0.046	0.036
California	-0.05	0.05	0.05	0.012	-0.043	-0.033
South Central	0.05	0.05	0.05	-0.035	0.007	-0.024
South Central	-0.05	-0.05	-0.05	0.038	-0.007	0.026
South Central	0.05	-0.05	-0.05	-0.005	-0.005	0.043
South Central	-0.05	0.05	0.05	0.006	0.005	-0.040

There was an obvious inverse effect on the change in prices to customer demand: as price goes up, volume decreases and vice versa. We see that 4225 has the highest volume reaction to the price changes in both regions. In both regions, a change in bags prices generally shows an inverse effect on the other products; however, the volume of bags do not fluctuate as significantly.People in South Central show less of reaction to price changes overall compared to California, although Californians react less to price changes of bags.

Summary and Recommendations

As we observed in the EDA, the price and demand of avocados in the U.S. has an intuitive inverse trend and the highest sales are during the first quarter of the year.

After finishing our analysis we found that overall Bags, which has the highest market share, had the lowest price elasticities in all regions. People didn't react highly on its price changes, which may indicate that this is the favorite avocado product, especially in

California. Product 4046's price elasticities varied across the regions. Product 4225 is the most elastic in all regions.

We recommend not increasing 4225's price to avoid decreased demand. Also, an increase in the price of bags could improve profit given its minimal change in volume while also increasing volume in other categories. Lastly, sellers can focus more on increasing their prices in the South Central region because customers don't change buying behavior significantly.

One improvement could be done in our project: we had to manually type in the unit price data from the website which resulted in a limited dataset. If we can get more data points, the results will be much more significant for our elasticity model. Also, there was little variation among the different prices of avocados, so the elasticities were not as significant as they could be. In the future, we could find data on avocado products that are more different and have larger price variations.