Elasticity Analysis on Avocado Supply and Demand

Hanchun Jiang

2021/11/10

```
## New names:
## * `` -> ...1
```

```
## Rows: 18249 Columns: 14
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): type, region
## dbl  (11): ...1, AveragePrice, Total Volume, 4046, 4225, 4770, Total Bags, S...
## date  (1): Date
```

```
##
## 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.
```

*The codes can be found at: https://github.com/carolinejiang757/Elasticity-Analysis-of-Avocado-Supply-and-Demand

1. Data Cleaning

Do any data cleaning you think appropriate on the data set and report any changes to the data that you make and why you made those changes.

There are two main aspects of the raw data set that can be improved: region and type. The original "region" column contains regions of different levels, and the "type" column prevents me from analyzing the two types together. Therefore, I made the following changes and made new data sets:

avocado:

The original data set. A column "year" is added to store the year value of a particular data entry.

US:

As the region values in the "region" column contain regions of different level, I made the "US" data set specific to data with region value "TotalUS".

wei_avocado:

As there is a type column in the original data frame, I made a weighted full data frame where "volume" is the total volume of different types, and "price" is the weighted price over two types.

wei_US:

As the region values in the "region" column contain regions of different level, I broke "wei_avocado" into smaller data frames according to region level. "wei_US" is the data frame for weighted national data.

wei_big_region:

This data set contains weighted data of big regions (defined as regions larger than a state).

wei_small_region:

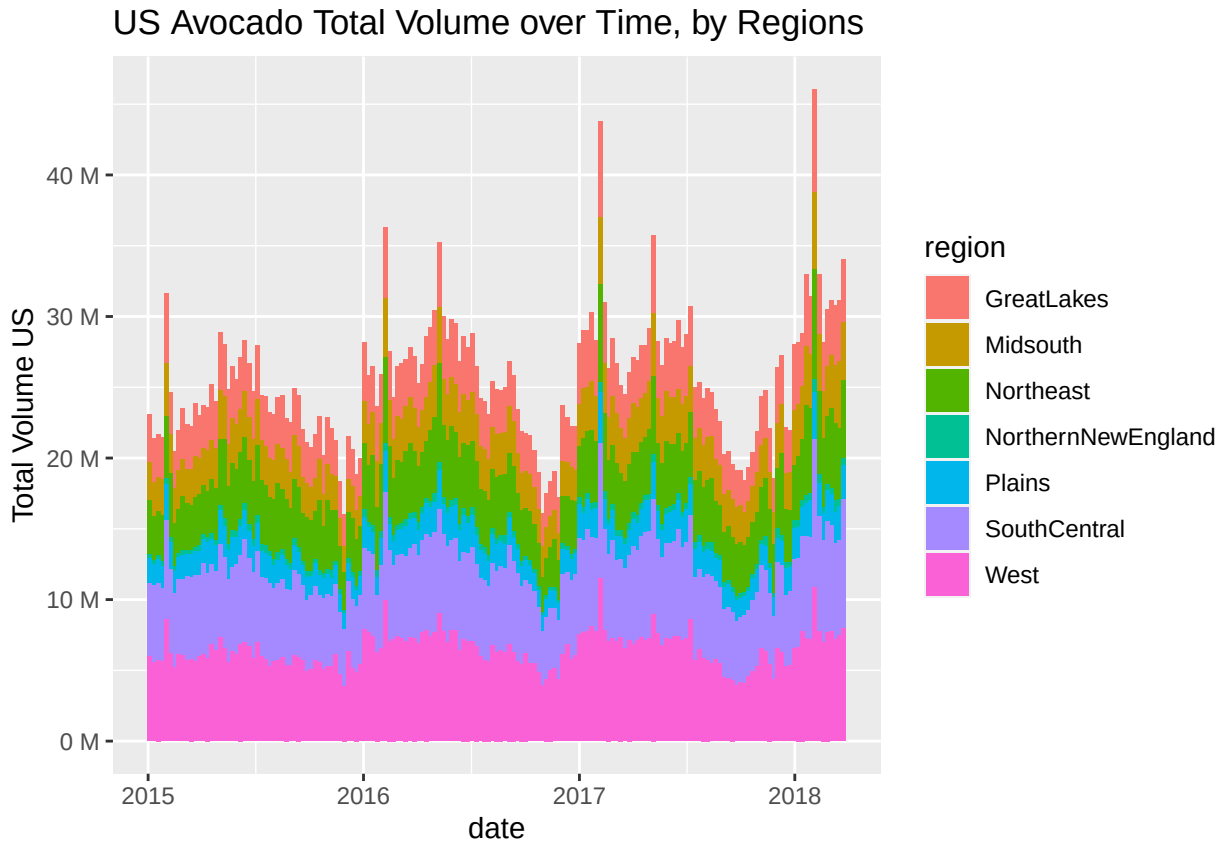
This data set contains weighted data of small regions (defined as regions smaller than or equal to a state).

2. Three Figures

Create three figures that collectively do the best job of describing and summarizing the data, i.e., after seeing your three figures, I should say, "ok, now i understand the basic patterns in the avocado market."

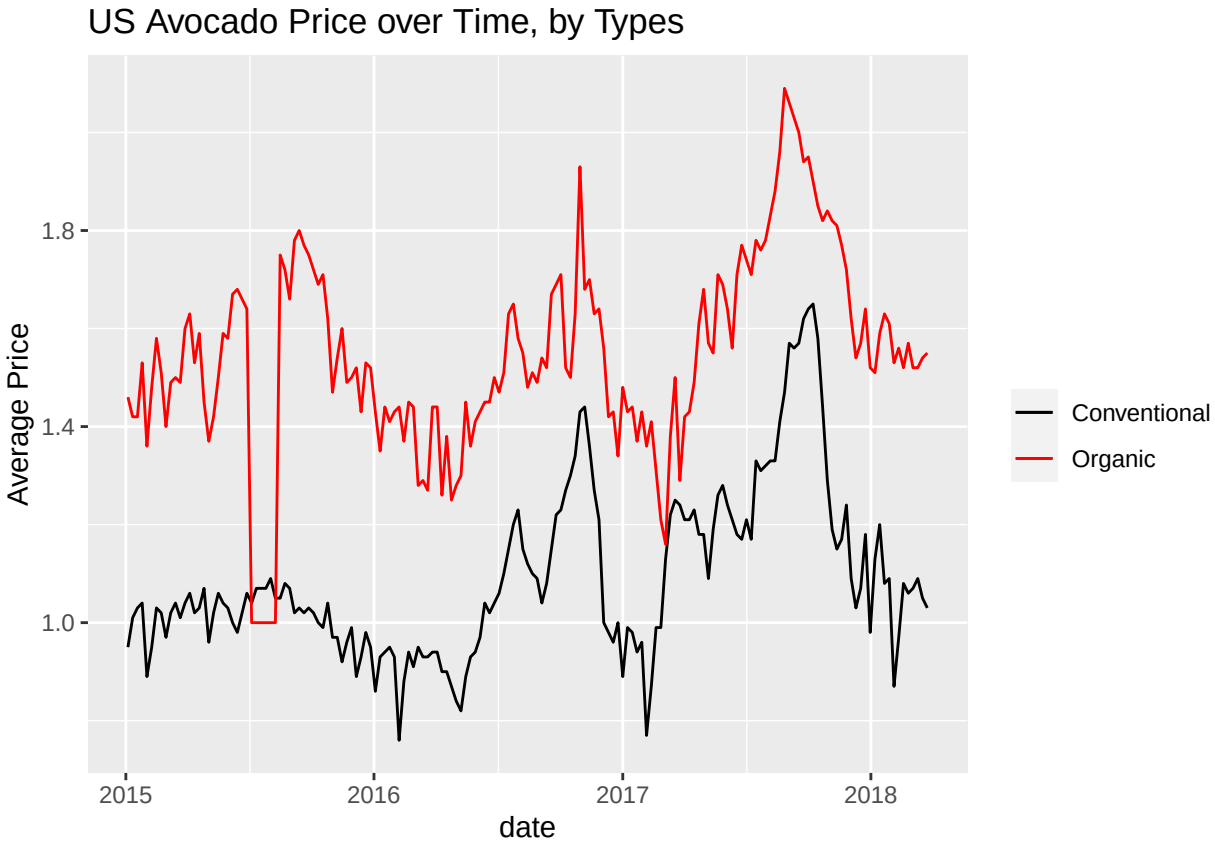
1) Volume, Region and Time

This figure shows how the total volume of avocado consumed in the US fluctuates over time. It also shows the composition of the total volume by regions, given a certain time.



2) Price, Type, and Time

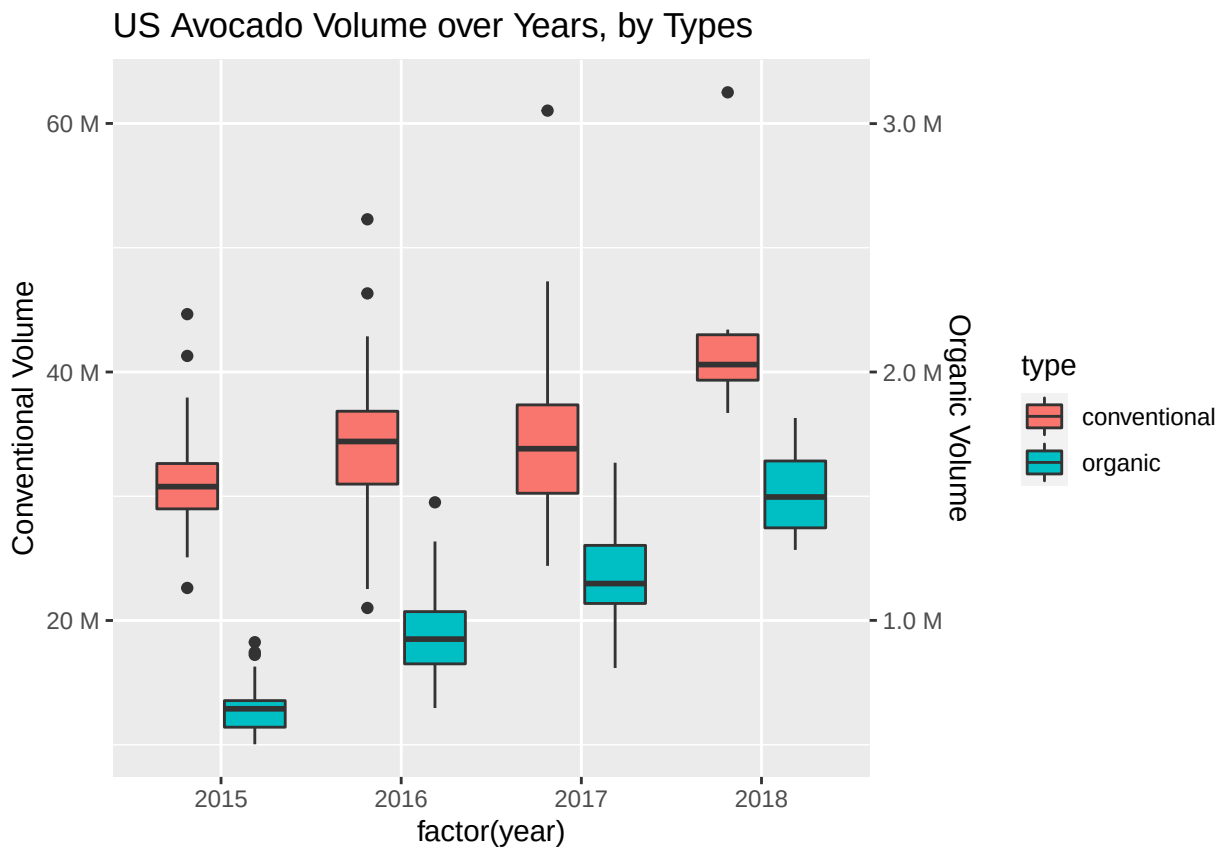
This figure shows how the price of conventional and organic avocado in the US fluctuates over time. It also shows the comparison of conventional and organic avocado prices, given a certain time.



3) Volume, Type and Time

This figure shows how the volume of conventional and organic avocado in the US fluctuates over the four years.

*Please note the conventional type corresponds to the left Y axis, and the organic type corresponds to the right Y axis.



3. Run Regressions

Run 1) time series, 2) cross-section, and 3) panel data regressions with the log quantity of avocados as the left hand side variable and the log price of avocados as the right hand side variable. Don't include any extra control variables, other than year dummies and geographic dummies if appropriate for the particular regression you are running. Discuss the coefficients you obtain. How would you describe what those coefficients are in words? How do they relate to economic parameters people care about? Would you say the time series variation and the cross-sectional variation are more driven by fluctuations in supply factors or demand factors?

1) Time Series

Time series data frame is the "wei_US" data frame created in Q1. It shows how price and volume of avocado change over time in the US.

Time series data frame:

```
## # A tibble: 169 x 5
## # Groups:   region, date [169]
##    region  date          volume price  year
##    <chr>   <date>         <dbl> <dbl> <dbl>
##  1 TotalUS 2015-01-04 31937188. 0.960  2015
##  2 TotalUS 2015-01-11 29733072. 1.02   2015
##  3 TotalUS 2015-01-18 29756579. 1.04   2015
##  4 TotalUS 2015-01-25 29026680. 1.05   2015
```

```
##  5 TotalUS 2015-02-01 45396358. 0.898  2015
##  6 TotalUS 2015-02-08 32868207. 0.962  2015
##  7 TotalUS 2015-02-15 28628698. 1.04   2015
##  8 TotalUS 2015-02-22 30610176. 1.03   2015
##  9 TotalUS 2015-03-01 33808499. 0.980  2015
## 10 TotalUS 2015-03-08 30878612. 1.03   2015
## # ... with 159 more rows
```

Time series regression:

```
##
## Call:
## lm(formula = log(volume) ~ log(price) + factor(year), data = wei_US)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33112 -0.06404 -0.00156  0.06758  0.25211
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      17.28864    0.01478 1169.713  < 2e-16 ***
## log(price)       -0.88322    0.06337  -13.936  < 2e-16 ***
## factor(year)2016  0.10738    0.02087    5.145 7.58e-07 ***
## factor(year)2017  0.24737    0.02356   10.502  < 2e-16 ***
## factor(year)2018  0.35674    0.03414   10.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1061 on 164 degrees of freedom
## Multiple R-squared:  0.6322, Adjusted R-squared:  0.6232
## F-statistic: 70.46 on 4 and 164 DF,  p-value: < 2.2e-16
```

2) Cross Sectional

Cross sectional data frame is created as the follows. It compresses the time dimension using average values, and shows how price and volume of avocado differ in different small regions.

Cross sectional data frame:

```
## # A tibble: 46 x 3
##    region                volume price
##    <chr>                  <dbl> <dbl>
##  1 Albany                95076.  1.36
##  2 Atlanta              524291.  1.08
##  3 BaltimoreWashington  797124.  1.35
##  4 Boise                 85285.  1.09
##  5 Boston               575586.  1.31
##  6 BuffaloRochester     135873.  1.39
##  7 California          6088649.  1.12
##  8 Charlotte            210388.  1.29
##  9 Chicago              791138.  1.38
## 10 CincinnatiDayton     263444.  1.03
## # ... with 36 more rows
```

Cross sectional regression:

```
##
## Call:
## lm(formula = log(volume) ~ log(price), data = cross)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63073 -0.67960 -0.03806  0.54392  2.69095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.1497     0.2046  64.280   <2e-16 ***
## log(price)   -1.9432     0.9725  -1.998   0.0519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9773 on 44 degrees of freedom
## Multiple R-squared:  0.08318,    Adjusted R-squared:  0.06235
## F-statistic: 3.992 on 1 and 44 DF,  p-value: 0.05192
```

3) Panel Data

Panel data frame is the "wei_small_region" data frame created in Q1. It shows how price and volume of avocado change with two variables: time and region.

Panel data frame:

```
## # A tibble: 7,772 x 5
## # Groups:   region, date [7,772]
##    region date       volume price  year
##    <chr>  <date>      <dbl> <dbl> <dbl>
##  1 Albany 2015-01-04 42247.  1.24  2015
##  2 Albany 2015-01-11 42378.  1.25  2015
##  3 Albany 2015-01-18 45630.  1.19  2015
##  4 Albany 2015-01-25 46263.  1.08  2015
##  5 Albany 2015-02-01 72102.  1.00  2015
##  6 Albany 2015-02-08 53025.  1.01  2015
##  7 Albany 2015-02-15 42750.  1.08  2015
##  8 Albany 2015-02-22 46827.  1.09  2015
##  9 Albany 2015-03-01 57259.  1.01  2015
## 10 Albany 2015-03-08 42134.  1.10  2015
## # ... with 7,762 more rows
```

Panel data regression:

```
##
## Call:
## lm(formula = log(volume) ~ log(price) + year + region, data = wei_small_region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79143 -0.09215  0.00877  0.09679  0.69675
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.641e+02  3.782e+00 -69.830  < 2e-16 ***
```

```
## log(price)                -9.143e-01  9.968e-03 -91.723  < 2e-16 ***
## year                       1.368e-01  1.876e-03  72.903  < 2e-16 ***
## regionAtlanta              1.523e+00  1.633e-02  93.248  < 2e-16 ***
## regionBaltimoreWashington  2.158e+00  1.616e-02 133.503  < 2e-16 ***
## regionBoise               -2.993e-01  1.633e-02 -18.329  < 2e-16 ***
## regionBoston               1.800e+00  1.617e-02 111.344  < 2e-16 ***
## regionBuffaloRochester     4.025e-01  1.617e-02  24.897  < 2e-16 ***
## regionCalifornia           4.007e+00  1.629e-02 246.011  < 2e-16 ***
## regionCharlotte            7.789e-01  1.617e-02  48.163  < 2e-16 ***
## regionChicago              2.150e+00  1.616e-02 133.015  < 2e-16 ***
## regionCincinnatiDayton     7.883e-01  1.641e-02  48.046  < 2e-16 ***
## regionColumbus             4.304e-01  1.633e-02  26.357  < 2e-16 ***
## regionDallasFtWorth        2.175e+00  1.681e-02 129.379  < 2e-16 ***
## regionDenver               1.972e+00  1.633e-02 120.755  < 2e-16 ***
## regionDetroit              1.215e+00  1.627e-02  74.645  < 2e-16 ***
## regionGrandRapids          5.949e-01  1.617e-02  36.790  < 2e-16 ***
## regionHarrisburgScranton   9.348e-01  1.617e-02  57.791  < 2e-16 ***
## regionHartfordSpringfield  1.224e+00  1.617e-02  75.682  < 2e-16 ***
## regionHouston              2.115e+00  1.689e-02 125.190  < 2e-16 ***
## regionIndianapolis         5.048e-01  1.625e-02  31.063  < 2e-16 ***
## regionJacksonville         4.700e-01  1.622e-02  28.986  < 2e-16 ***
## regionLasVegas             9.928e-01  1.640e-02  60.550  < 2e-16 ***
## regionLosAngeles           3.177e+00  1.650e-02 192.527  < 2e-16 ***
## regionLouisville          -1.604e-01  1.629e-02  -9.846  < 2e-16 ***
## regionMiamiFtLauderdale    1.731e+00  1.619e-02 106.954  < 2e-16 ***
## regionNashville            5.555e-01  1.641e-02  33.859  < 2e-16 ***
## regionNewOrleansMobile     8.633e-01  1.632e-02  52.905  < 2e-16 ***
## regionNewYork              2.767e+00  1.617e-02 171.145  < 2e-16 ***
## regionOrlando              1.201e+00  1.620e-02  74.098  < 2e-16 ***
## regionPhiladelphia         1.565e+00  1.617e-02  96.767  < 2e-16 ***
## regionPhoenixTucson        1.952e+00  1.731e-02 112.756  < 2e-16 ***
## regionPittsburgh           9.697e-02  1.618e-02   5.994 2.14e-09 ***
## regionPortland             1.712e+00  1.636e-02 104.675  < 2e-16 ***
## regionRaleighGreensboro    1.055e+00  1.619e-02  65.155  < 2e-16 ***
## regionRichmondNorfolk      8.319e-01  1.626e-02  51.156  < 2e-16 ***
## regionRoanoke              2.889e-01  1.629e-02  17.742  < 2e-16 ***
## regionSacramento           1.520e+00  1.617e-02  93.971  < 2e-16 ***
## regionSanDiego             1.527e+00  1.634e-02  93.450  < 2e-16 ***
## regionSanFrancisco         2.190e+00  1.617e-02 135.441  < 2e-16 ***
## regionSeattle              1.808e+00  1.622e-02 111.443  < 2e-16 ***
## regionSouthCarolina        1.209e+00  1.624e-02  74.425  < 2e-16 ***
## regionSoutheast            3.524e+00  1.623e-02 217.064  < 2e-16 ***
## regionSpokane             -1.774e-01  1.627e-02 -10.905  < 2e-16 ***
## regionStLouis              5.922e-01  1.622e-02  36.523  < 2e-16 ***
## regionSyracuse            -3.348e-01  1.617e-02 -20.709  < 2e-16 ***
## regionTampa                1.303e+00  1.621e-02  80.392  < 2e-16 ***
## regionWestTexNewMexico     1.805e+00  1.685e-02 107.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1486 on 7724 degrees of freedom
## Multiple R-squared:  0.9792, Adjusted R-squared:  0.9791
## F-statistic:  7739 on 47 and 7724 DF,  p-value: < 2.2e-16
```

4) Interpretation

Coefficient for log(price) means how much log(volume) changes when log(price) changes in one unit. This shows elasticity of the volume of avocado over price change, though it's yet not clear whether this is supply or demand elasticity.

Coefficient for the years means how log(volume) differs when in different years. This shows how volumes of avocados are inherently different between different years, due to varied demand and supply over the years.

Coefficient for the regions means how log(volume) differs when in different regions. This shows how volumes of avocados are inherently different between different regions, due to varied demand and supply between different regions.

I think time series variation is more driven by supply. Looking at the graph in Q4, which illustrates time series variation of volume and price, I find that most of the times volume and price are moving in opposite directions, indicating the impact of supply (more supply, lower price; less supply, higher price).

Cross sectional variation, on the other hand, is more driven by demand. Intuitively, besides regions where avocados are harvested, the amount of avocodos transported and sold in each city is mainly determined by the prediction of the demand for avocado in that city.

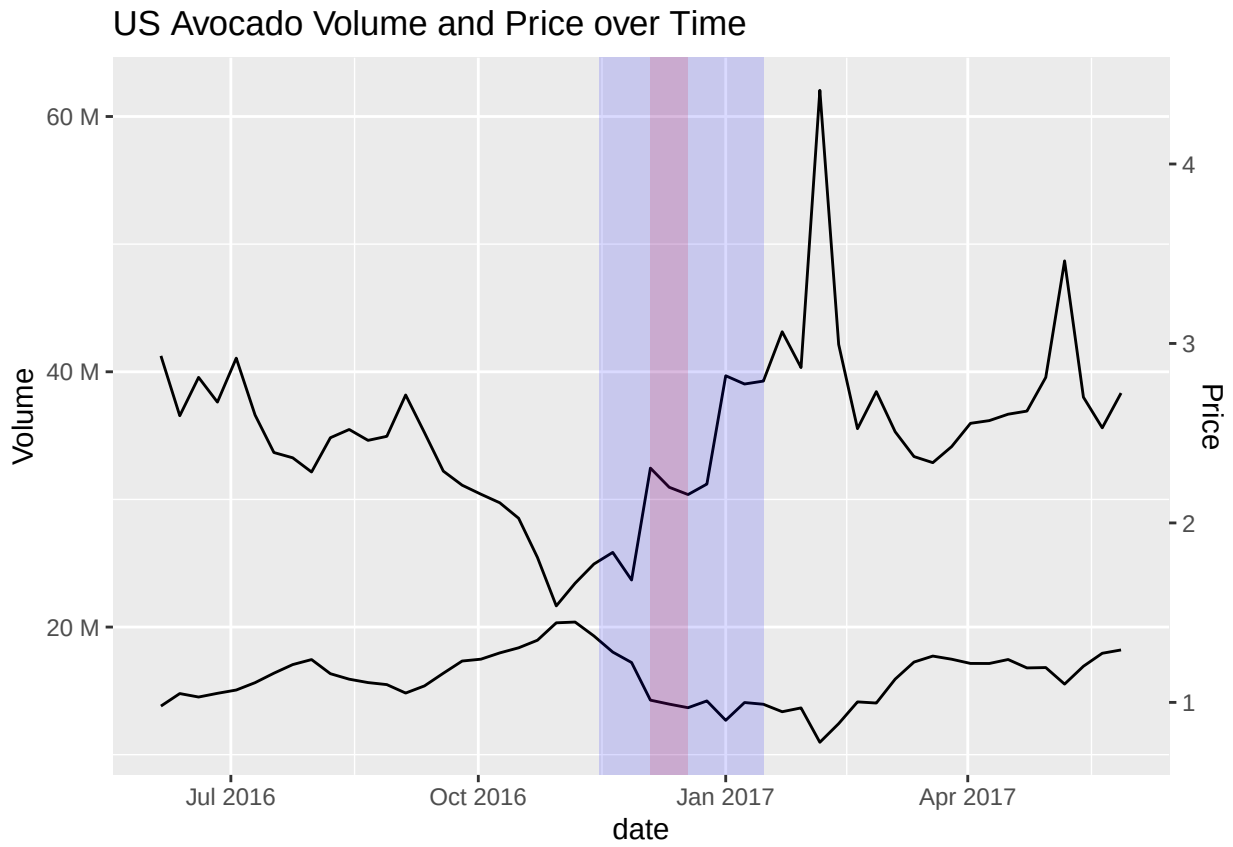
4&6. Elasticity of Supply and Graph

The department of agriculture wants your best estimate of the elasticity of supply. Run whatever regression you think best gets at that.

Explain why you made these choices. What estimate would you give for the elasticity of supply? How confident are you in your estimate, i.e. what standard error band would you put around your estimate? Make a graph that best explains your finding.

In the following graph, I found an area where volume and price are both decreasing (red area). As volume and price are moving in the same directions, this indicates a demand shock. Intuitively, this period (2016-12-4 to 2016-12-18) is the period when people are preparing for the Christmas, and as avocado is not among the traditional ingredients for Christmas, and people tend to buy the its substitute goods, the demand for avocado decreased, which caused a demand shock.

```
## [1] "English_United States.1252"
```



By running regression over the two months surrounding this demand shock (blue area), and making a dummy variable which is "0" when not during the shock and "1" when during the shock, I found out the change of volume due to the shock by looking at the coefficient for the dummy. The change of volume is -0.13540. Please note that this coefficient is not very significant with a p value of 0.1, because this is a slight demand shock, which caused a slight change in price and lasted for a short period of time. However, this demand shock is already the most significant one compared to other periods of time, as it is the longest time span when volume and price are moving in the same direction.

```
##
## Call:
## lm(formula = log(volume) ~ log(price) + shock, data = supply_elas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11120 -0.05952 -0.01058  0.07010  0.09978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.38208    0.04271 407.025 1.48e-14 ***
## log(price)  -1.50289    0.30954  -4.855  0.00284 **
## shock1      -0.13540    0.07004  -1.933  0.10139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09481 on 6 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:  0.7322
## F-statistic: 11.94 on 2 and 6 DF,  p-value: 0.008099
```

I also need to find out the change of price caused by the shock. The average price not during the shock minus the average price during the shock is 0.075. Deviding the change of volume by the change of price, I get the supply elasticity, which is 0.557.

```
## [1] 0.5571509
```

The standard error given by the regression is 0.07. The actual standard error should be higher due to many factors in reality not being ideal (such as measurement error), and also due to the fact that the coefficient is divided by another estimate (i.e. estimate for the change in price). I will report a 0.2 standard error.

5&6. Elasticity of Demand and Graph

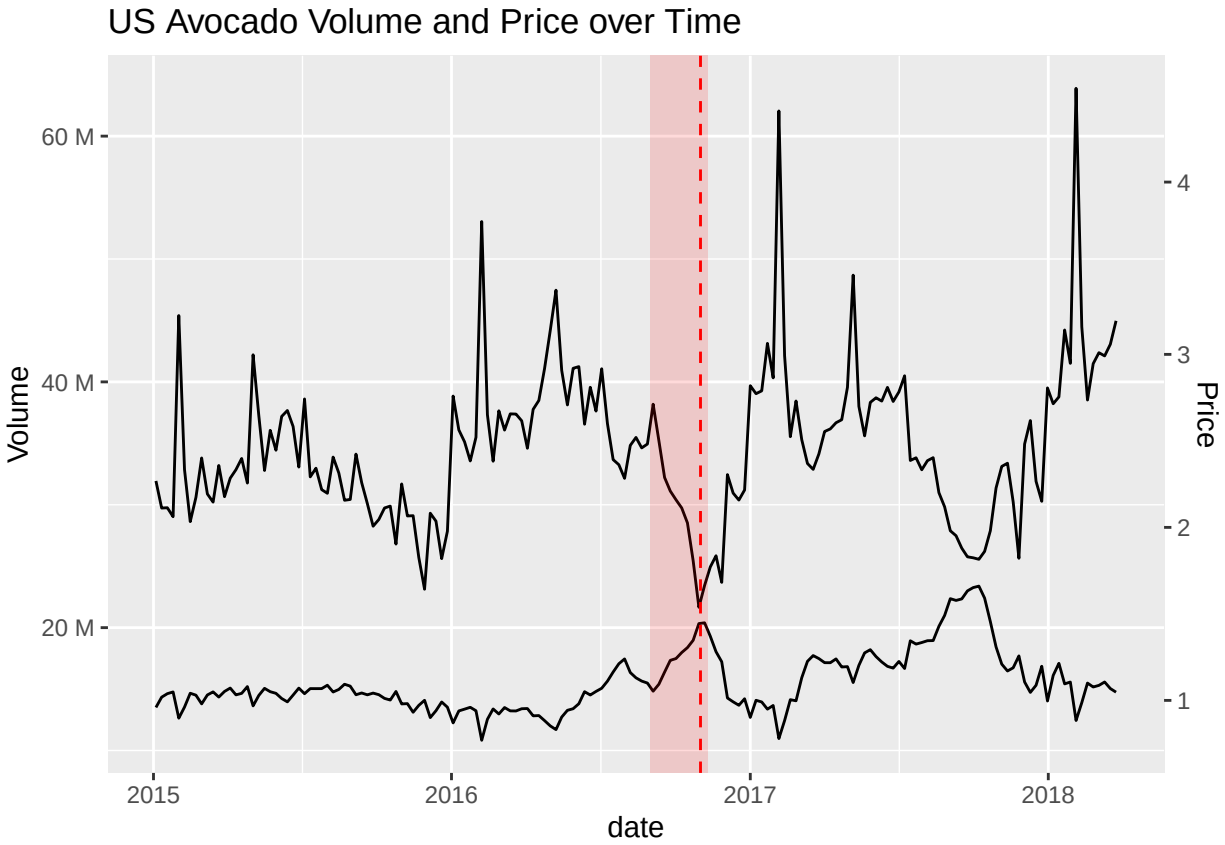
The department of agriculture wants your best estimate of the elasticity of demand. Run whatever regression you think best gets at that.

Explain why you made these choices. What estimate would you give for the elasticity of demand? How confident are you in your estimate, i.e. what standard error band would you put around your estimate? Make a graph that best explains your finding.

In the following graph, I marked a drastic volume decrease and corresponding price increase with the red line and the shaded area. As volume and price are moving in the opposite directions, this indicates a supply shock.

The news below also confirms that the marked area indicates a supply shock in around November, 2016 caused by a strike in Mexico and a drought in California.

<https://www.forbes.com/sites/geoffwilliams/2016/10/31/how-the-avocado-shortage-is-affecting-chipotle-grocers-and-you/?sh=7344b52b5475>



By running regression on this supply shock period, I found the demand elasticity over price is -1.60877.

```
##
## Call:
## lm(formula = log(volume) ~ log(price), data = demand_elas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07140 -0.02190 -0.00008  0.03228  0.04731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.55441    0.03141  558.87  < 2e-16 ***
## log(price)  -1.60877    0.12635  -12.73 1.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04057 on 8 degrees of freedom
## Multiple R-squared:  0.953,  Adjusted R-squared:  0.9471
## F-statistic: 162.1 on 1 and 8 DF,  p-value: 1.363e-06
```

The standard error given by the regression is 0.12635. The actual standard error should be higher due to many factors in reality not being ideal, such as measurement error. I will report a 0.2 standard error.

7. Advice on Data Collection

As the most difficult and time consuming part during the data analysis process is distinguishing the impact of supply from the impact of demand, the Department of Agriculture should perform data collection in ways that distinguish supply from demand.

Advice: For total US data, the Department of Agriculture should collect data on how much avocados are produced, imported and exported in the US. For regional data, they should collect data on how much avocados are produced, transported into the region, and exported from the region. In this way, the supply of avocado could be calculated, and it will be easier to perform analysis related to supply and demand.