Corey McCrea

CPSC 298 – Intro to R

5/14/19

Avocado Analysis

Link to Slides: https://bit.ly/presentMe

Dataset/Summary

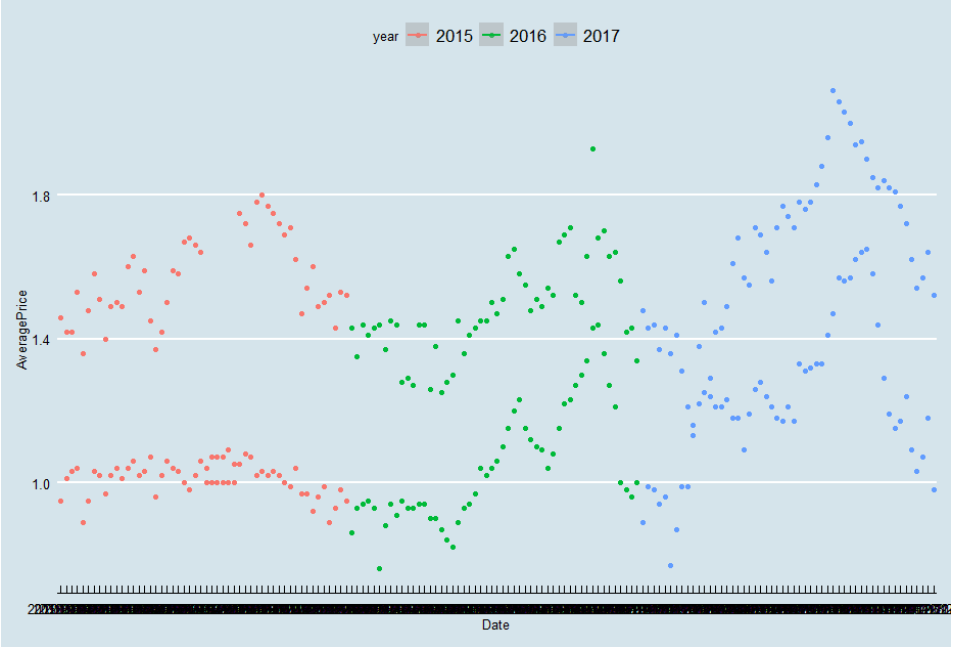
I found a Kaggle dataset on avocado pricing and volume over the years 2015-2018. It provides 12 stats (date, average price, total volume, type, year, region) regarding avocados over a three-year period (2015-2018). Date is the weekly date of observation. Average price is the average price of a single avocado that week. Type is whether the avocado is conventional or organic. Year is the year of observation. Region is the city or region that the avocados are being sold. Total volume is the number of avocados sold. 4046, 4225, 4770 are the different avocado types, i.e. small Haas etc.

In order to use this dataset effectively, I had to do some cleaning first. I only looked at the total US region and I ignored 2018 as there wasn't a complete year of entries.

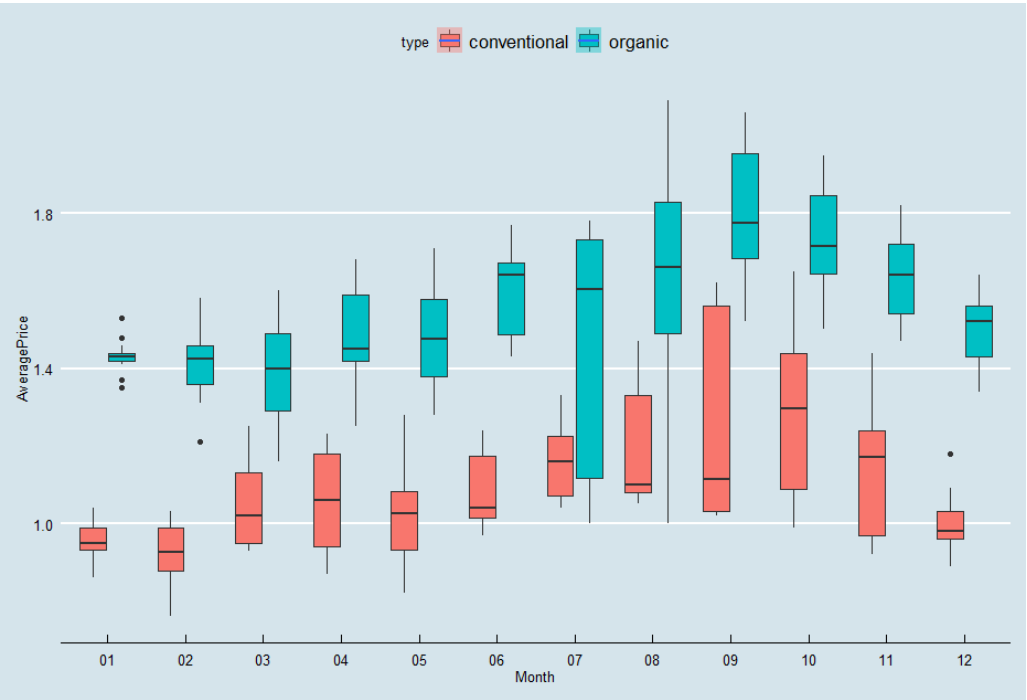
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year
2	51	2015-01-04	0.95	31324277.73	12357161.34	13624083.05	844093.32	4498940.02	3585321.58	894945.63	18672.81	conventional	2015
3	51	2015-01-04	1.46	612910.15	233286.13	216611.2	4370.99	158641.83	115068.71	43573.12	0	organic	2015
4	50	2015-01-11	1.01	29063542.75	11544810.53	12134773.38	866574.66	4517384.18	3783261.16	718333.87	15789.15	conventional	2015
5	50	2015-01-11	1.42	669528.88	270966.74	260971.6	3830.42	133760.12	106844.49	26915.63	0	organic	2015
6	49	2015-01-18	1.03	29043458.85	11858139.34	11701947.8	831301.9	4652069.81	3873041.26	771093.2	7935.35	conventional	2015
7	49	2015-01-18	1.42	713120	254319.58	311811.01	4020.85	142968.56	101850.23	41118.33	0	organic	2015
8	48	2015-01-25	1.04	28470310.84	12167445.03	10734652.82	768020.05	4800192.94	3978636.9	812924.73	8631.31	conventional	2015
9	48	2015-01-25	1.53	556368.86	207494.87	212312.02	4753.87	131808.1	95964.83	35843.27	0	organic	2015
10	47	2015-02-01	0.89	44655461.51	18933038.04	18956479.74	1381516.11	5384427.62	4216452.03	1121076.47	46899.12	conventional	2015
11	47	2015-02-01	1.36	740896.97	302561.47	259286.44	5852.28	173196.78	129953.15	43243.63	0	organic	2015
12	46	2015-02-08	0.95	32137333.01	13308193.4	13381347.54	737939.45	4709852.62	4022474.85	673453.54	13924.23	conventional	2015
13	46	2015-02-08	1.48	730874.31	215657.99	273897.84	5316.55	236001.93	179887.47	56114.46	0	organic	2015
14	45	2015-02-15	1.03	28012520.93	12626615.3	9783489.59	845653.52	4756762.52	4096226.46	648632	11904.06	conventional	2015
15	45	2015-02-15	1.58	616177	207650.64	228653.33	4950.01	174923.02	140602	34321.02	0	organic	2015
16	44	2015-02-22	1.02	29936729.76	12628562.36	11354281.64	937138.85	5016746.91	4336247.12	667149.29	13350.5	conventional	2015
17	44	2015-02-22	1.51	673446.69	272594.88	230342.87	4883.55	165625.39	133314.63	32310.76	0	organic	2015
18	43	2015-03-01	0.97	32994014.16	13282222.98	13733124.48	1070576.07	4908090.63	4129138.63	725218.35	53733.65	conventional	2015
19	43	2015-03-01	1.4	814484.79	361996.84	275458.57	5556.02	171473.36	148488.14	22985.22	0	organic	2015
20	42	2015-03-08	1.02	30094698.85	13013750.35	10973972.6	834009.15	5272966.75	4583726.82	673149.42	16090.51	conventional	2015
21	42	2015-03-08	1.49	783913.05	253078.17	327972.28	5658.7	197203.9	169302.64	27901.26	0	organic	2015
22	41	2015-03-15	1.04	29572225.71	13149988.71	10634070.76	871575.04	4916591.2	4287621.56	614904.92	14064.72	conventional	2015
23	41	2015-03-15	1.5	644584.67	235569.01	258475.58	5308.79	145231.29	109325.32	35905.97	0	organic	2015
24	40	2015-03-22	1.01	32513550.51	13697405.61	12659784.83	1066385.92	5089974.15	4275071.46	762527.54	52375.15	conventional	2015
25	40	2015-03-22	1.49	682640.03	237139.46	320353.82	5290.59	119856.16	83554.82	36301.34	0	organic	2015
26	39	2015-03-29	1.04	29982648.43	12524637.04	11541041.35	811272.88	5105697.16	4487886.76	610349.2	7461.2	conventional	2015
27	39	2015-03-29	1.6	674551.02	193273.88	324919.62	5794.39	150563.13	113703.56	36859.57	0	organic	2015
28	38	2015-04-05	1.06	31500669.44	13939014.43	11526980.36	871981.29	5162693.36	4477299.71	666514.74	18878.91	conventional	2015
29	38	2015-04-05	1.63	661842.02	197746.89	275595.72	6434.88	182064.53	131692.64	50371.89	0	organic	2015
30	37	2015-04-12	1.02	32046401.64	14793354.18	11210544.11	807942.41	5234560.94	4400469.38	826567.05	7524.51	conventional	2015

Analysis/Graphs

The first thing I looked at was average price and aate to get a general, overarching idea of the trends in this dataset.



Here a cyclical pattern can be seen going on throughout the years, which could be due to avocadoes going in season during the summer and then out of season. There is a spike in 2017 which could be explained as a result of the California fires impacting avocado groves as California is one of the largest avocado suppliers. We also see two groups the upper group, organic and the lower group conventional. This can be seen more clearly in this average price and month box plot; this contains data from all three years.

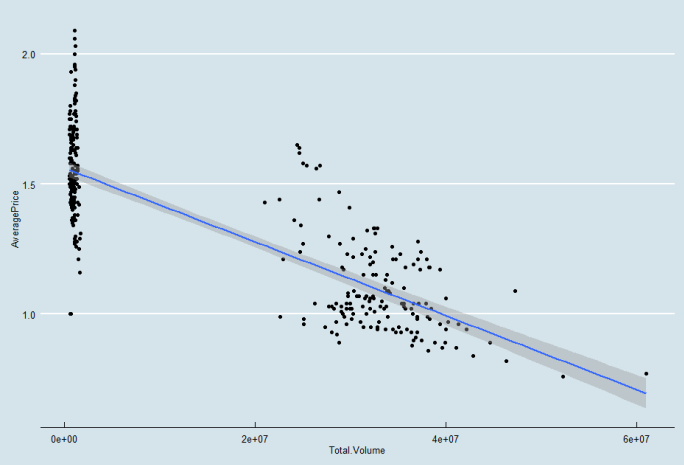


I sought to confirm this relationship with an **ANOVA** analyzing average price with month combined with type.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	3.959	0.360	13.65	<2e-16
type	1	15.964	15.964	605.32	<2e-16
Residuals	301	7.938	0.026		

There was a significant relationship between these variables as our p-value is < .05. This is evident from the above boxplot.

I then ran a **regression model** to explore the expected relationship between price and volume.

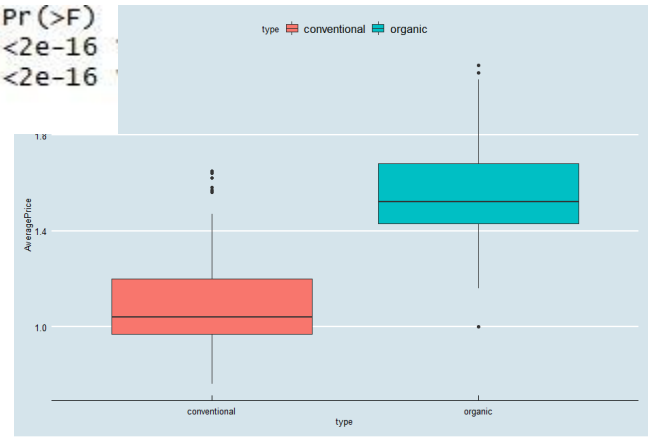



```
Call:
lm(formula = AveragePrice ~ Total.Volume, data = avo)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5530 -0.1173 -0.0336  0.1274  0.5437

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.561e+00  1.487e-02  104.96   <2e-16 ***
Total.Volume -1.418e-08  6.266e-10  -22.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1839 on 312 degrees of freedom
Multiple R-squared:  0.6215,    Adjusted R-squared:  0.6203
F-statistic: 512.2 on 1 and 312 DF,  p-value: < 2.2e-16
```

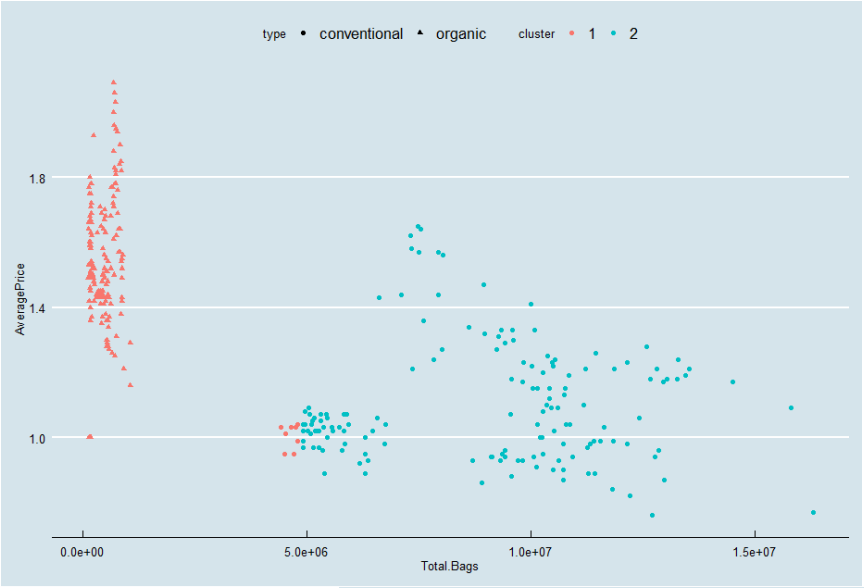
Due to a p-value of <.05 there is a significant relationship between price and volume as one would expect. It is clearly a negative relationship as is seen in the graph, as volume goes up the price goes down. From a business standpoint this is very logical, supply and demand are directly related.

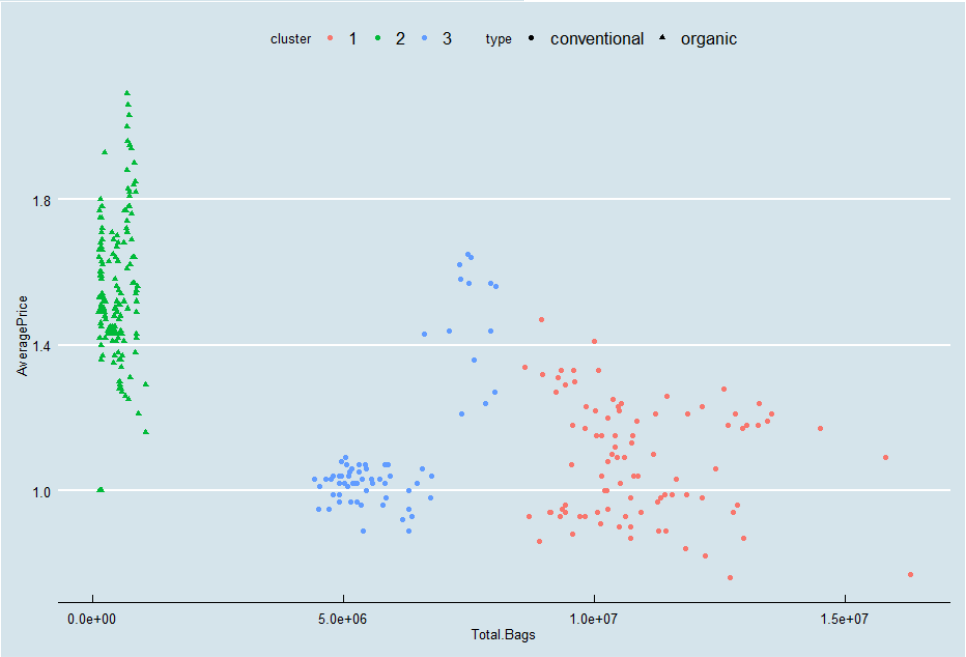
I then wanted to confirm that price and type of avocado were related so I ran another **ANOVA**. I found that there is a significant relationship between these two. Organic avocado prices average a whole 50 cents higher than conventional avocadoes.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	1122.2	1122.2	11811.5	<2e-16
year	3	103.3	34.4	362.6	<2e-16
Residuals	18244	1733.3	0.1		



For my fourth analysis I ran a **k-means clustering** starting with k =2 to cluster sales into two groups based on total bags and average price. The clustered groups were very close to being separated by type, so I then increase the k value to three and one group was organic and two were conventional. If I had more time I would have enjoyed exploring if there was a relationship between the two conventional groups.





Finally, I created a **logistic regression** model to predict whether an avocado was organic or conventional solely using average price. This model performed very well (~90% accuracy) considering it only takes one factor into account. This could be improved by adding more relevant variables.

	FALSE	TRUE
conventional	143	14
organic	18	139