I wanted to see if the total volume and average price of avocados could be modeled from some of the other columns in this dataset. I believe ­­the Hass Avocado Board would be very interested to know this because pricing and demand are fundamental concepts in a business. Understanding pricing and demand will give ­­­­the Hass Avocado Board fundamental insight into their business at large. By analyzing the total volume of avocadoes sold, the supply chain can be optimized to keep up with demand and marketing dollars can be funneled towards market segments where demand is low. Also, by understanding what factors affect price, one could use that knowledge to optimize income and increase profitability.

**Investigating Total Volume**

I first wanted to see what kind of data may drive (or is primarily responsible for) the total volume of avocadoes sold.

Using Linear Regression

I first tried to see if Total Volume was related to any of the other predictor data in a linear fashion. Some of the columns in the dataset are actually more-specific breakdowns of Total Volume. There are three columns referring to PLU codes of avocadoes sold and there are three columns referring to the size of bags sold. It is likely these predictors would not be helpful in a logistic regression, but perhaps there are other predictors that would be helpful.



Given the requirements of the linear regression model, I technically had to model the log of Total Volume but the general relationships I found will still be valid on the actual Total Volume.

I created a large linear model using most of the predictors in the dataset. The fit seemed to match fairly well. As you can see in the figure, the error terms of the model had a fairly normal distribution.

The adjusted R-Squared value was about 0.73, however some of the predictors did not seem significant in the model. I suspected that the PLU and Bag predictors could be extremely collinear with the log of Total Volume. Therefore, I performed checks for collinearity

Using the vif function, I found that the PLU and Bag predictors were definitely collinear and could not be used in an accurate model. From the regsubsets function, I saw that 2-3 predictors would seem an ideal number for a final linear model (see figure to the right), but one of the ‘good’ predictors that regsubsets returned was Small.Bags.

Since that was clearly an unacceptable predictor from the vif function, I created a linear model using only 2 predictors (type and region) as predictors and applied it to a training subset of the data. I then tested the model on a validation subset. Again, the error terms from the model had an approximate normal distribution. The adjusted R-squared from the model was very good at approx. 0.96. The mean squared error after testing the model was also very low at approx 0.22.

Using Penalized Regression Methods

Penalized regression methods are usually helpful when many collinear variables exist in a dataset. This should make these techniques ideal for this dataset, but unfortunately, they did not yield useful results when looking at the Total Volume response.

Essentially, penalized regression methods will restrict (and sometimes eliminate) collinear predictors so that their variance is reduced. I ran a fairly large model across all three techniques; Ridge Regression, LASSO, and E-Net. I iterated slightly different models across these three methods and unfortunately, the collinear predictors remained while the potentially useful predictors were eliminated. Since the PLU and Bags columns perfectly predict Total Volume, the results were not helpful.

**🡪 Conclusions for Total Volume**

It appears that region and type are good predictors of logTotal.Volume in a linear regression model. We can be confident that further research into the type and region avocadoes sold will prove useful in understanding and managing the Total Volume sold.

**Investigating Average Price**

I also wanted to see what kind of data may drive (or is primarily responsible for) the average price of avocadoes sold.

Using Linear Regression

I also had to model the log of Average Price using linear regression; much like I did for Total Volume. I fit a large model using most of the predictors in the dataset. The Adjusted R-squared value was lackluster at approx. 0.61. The error terms still had an approximately normal distribution.

After running the vif and regsubsets functions, I again saw that the PLU and Bag predictors were definitely collinear. This time the regsubsets function indicated that 5 predictors should be ideal (see figure to the right). Many of the collinear predictors detected earlier were also shown to be some of the ‘good’ predictors to use. When I created a smaller linear model using the ‘best’ five predictors, the resulting Adjusted R-squared was approx. 0.60. The mean squared error on the test data was approx. 0.04. Although the MSE was quite low, the lackluster Adjusted R-squared measure on a ‘best’ model with several collinear predictors does not seem to be very useful.

One thing that is useful though is that two of the non-collinear ‘good’ predictors in this model are Type and Y ear. As we will see below using Ridge Regression, that method proved that these are the two truly best predictors for Average Price

Using Penalized Regression Methods

When I applied the penalized regression methods using Average Price as a response, the results were much more useful than for Total Volume. Each technique showed that type is the most significant predictor by far. Each year is influential as well, especially 2017. Every other predictor was reduced to almost zero. To the right is a graph of the ‘predictor space’ in the Ridge Regression model (that technique gave us slightly better performance out of all penalized methods). Each line represents a coefficient of the final model. The straight black line intercepts the colored lines where the coefficient values lie. The green and light blue lines represent the coefficients for Type and the Year 2017 respectively.

**🡪 Conclusions for Average Price**

It appears that Type and the Year of 2017 are good predictors of Average Price after running Ridge Regression. We can be confident that further research into the type and data from 2017 will prove useful in understanding what price Avocadoes sell for. There may have been conditions during 2017 that affected avocado prices. Such conditions would definitely be worth further research.

**Dataset Details**:

**“Avocado Prices”** <https://www.kaggle.com/neuromusic/avocado-prices/home>

Downloaded Dec 8, 2018.