Avocados Dataset from Kaggle

Ru Feng, Eva Zhang, Meredith Moran

Introduction

Our dataset consists of data from the Hass Avocado Board website. This dataset compiles avocado sale prices and volumes in many United States regions and cities from 2015 to 2018. There are 13 predictors and 18,249 observations in this set. Columns in the dataset include Date, AveragePrice, Type (conventional or organic), Year, Region, Total Volume (number of avocados sold), number of avocados sold for specific PLU codes 4046, 4225 and 4770, Total Bags, Small Bags, Large Bags and XLarge Bags.

We are interested in utilizing this dataset as it contains pertinent information to avocado sales. We hope to extract and convey important information that is useful and engaging to avocado lovers and non-lovers alike.

Questions

1) What factors have the greatest impact on the average price of an avocado?
2) Can we make predictions on future sales and prices of avocados given the data?
3) Can we guess the type of an avocado based on the various factors (pricing, sales volume, etc.)?

Analysis #1: Linear Regression

Initial fit: Linear regression gives us a sufficient idea of the relationship between all predictors and Average Price. First, we fit a linear regression based on all variables except the date predictor. The generated R-squared value is 0.558, and AIC is 3815. The R-squared value is not too low, indicating that there is a somewhat linear relationship between average price and all predictors. The predictors year, type (conventional or organic), and location yield the smallest p-values. Organic corresponds to a higher price as the coefficient for this predictor is positive, which is as expected. Also, there seems to be a heavy correlation between location and average price that we can further explore through subsequent linear regressions.

Using only year: The second linear regression model includes only the year predictor. We want to use linear regression to observe how average price behaves over time. However, linear regression has trouble handling the date predictor, which is a datetime object, and thus treated as a categorical variable. Due to this, there were a large number of dummy variables created (corresponding to each timestamp) that did not provide pertinent information. As a result, year is the only predictor used. For this model, there is an extremely low p-value corresponding to year. This confirms that there is a statistically significant effect of year on Average Price. However, year alone cannot accurately predict the Average Price given that the R-squared value is extremely low (R-squared = 0.009). (Figure 0a) This is another topic for further exploration, which is completed in the next section with trend fitting and forecasting.

Dep. Variable:	AveragePrice	**R-squared:**	0.009
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	159.9
Date:	Sun, 16 May 2021	**Prob (F-statistic):**	1.71e-36
Time:	20:23:47	**Log-Likelihood:**	-9214.4
No. Observations:	18249	**AIC:**	1.843e+04
Df Residuals:	18247	**BIC:**	1.845e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-79.0913	6.366	-12.423	0.000	-91.570	-66.613
year	0.0399	0.003	12.644	0.000	0.034	0.046

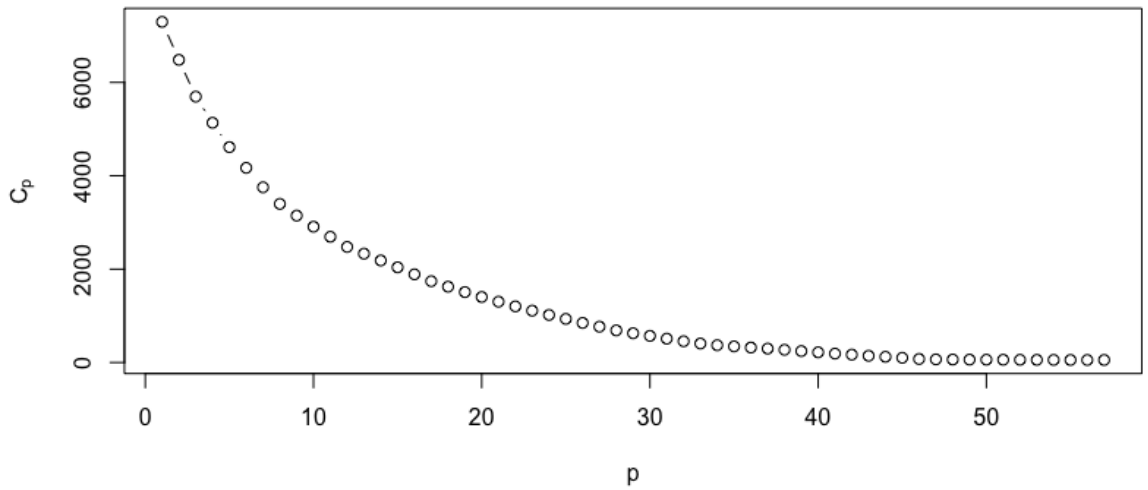
[Figure 0a: Linear Regression with only Year Summary]

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.1656	0.003	336.207	0.000	1.159	1.172
type[T.organic]	0.4881	0.005	102.466	0.000	0.479	0.497
TotalVolume	-3.008e-05	4.7e-05	-0.641	0.522	-0.000	6.2e-05
PLU4046	2.998e-05	4.7e-05	0.639	0.523	-6.2e-05	0.000
PLU4225	3.018e-05	4.7e-05	0.643	0.520	-6.18e-05	0.000
PLU4770	2.961e-05	4.7e-05	0.631	0.528	-6.24e-05	0.000
TotalBags	0.0050	0.036	0.139	0.889	-0.066	0.076
SmallBags	-0.0050	0.036	-0.139	0.890	-0.076	0.066
LargeBags	-0.0050	0.036	-0.139	0.890	-0.076	0.066
XLargeBags	-0.0050	0.036	-0.138	0.890	-0.076	0.066

[Figure 0b: Linear Regression w/o Date, Year, Region Summary]

Using all predictors but date, year, and region: The last linear regression model examines relationships of Average Price to all variables but date, year, and region. The p-values are again only small for intercept and type, as shown in Figure 0b above, consistent with the first regression. Also, the R-squared value for this model is 0.398, and AIC is 9344. Compared to the initial linear regression model, this model explains less variability, even taking into account the smaller number of predictors used indicated by the AIC value.

Variable selection: Evaluation using p-value alone is usually not accurate since it relies solely on the training set and is only robust when certain assumption criteria regarding the p-values are met. Variable selection provides a more robust idea of which variables are useful. The AIC variable selection above shows that every variable is important since the model that includes every predictor has the lowest AIC value. Additionally, a plot (Figure 0c) is produced in R for the number of predictors, p, and c_p to better show this relationship of model flexibility and model quality. We used the command “regsubsets” and the forward selection method. c_p is another model evaluation parameter that penalizes using too many predictors. We are using c_p instead of AIC because there was no easy way to generate this kind of plot using AIC. Furthermore, AIC and c_p are proportional, therefore they always select the same number of predictors. There is a proof for the proportionality included in the bibliography section at the end of this paper. A lower c_p value corresponds to a better model. In this case, a large number of predictors (p) yields a low c_p value. This result is somewhat expected, as sales of avocados should be related to every single variable.

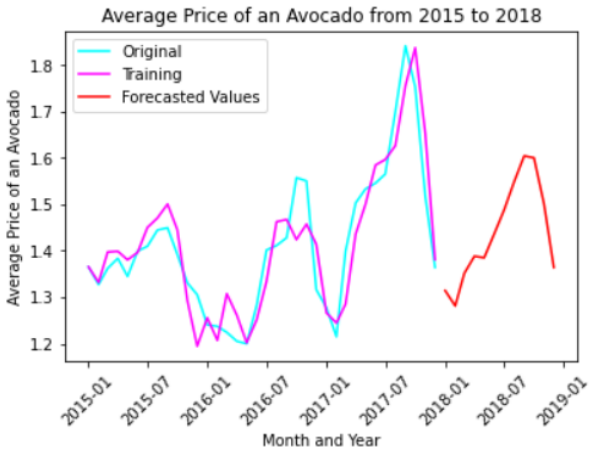


[Figure 0c: Number of Predictors vs. c_p (another model evaluation parameter, proportional to AIC)]

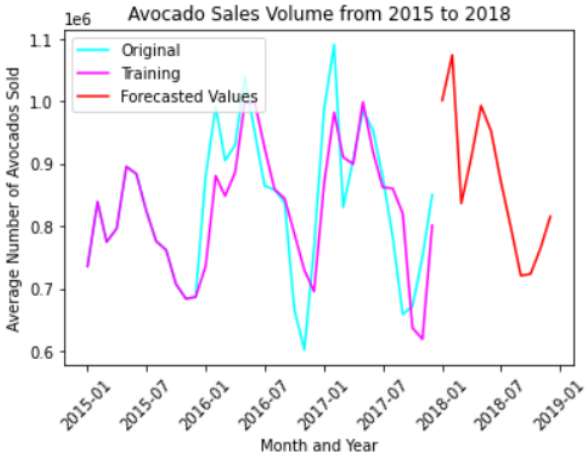
Subset Selection with Upper Bound for Number of Predictors: To decrease the number of predictors incorporated into the model, we restrict the maximum number of predictors the model can use by nvmax in regsubsets command in R: forward.fit ← regsubsets(AveragePrice~., data=avocado, nvmax=12, method="forward").
Using the max number of variables = 12, "regsubsets" command selects variables consistent with all of our previous observations. It selected predictors: type, year, and dummy variables for some different regions, where the regions are Charlotte, DallasFtWorth, HartfordSpringfield, Houston, NewYork, Northeast, Philadelphia, Sacramento, Sanfrancisco, and SouthCentral. Again, the selection of variables type and year here confirm our belief that these predictors have a significant effect on Average Price.

Mean Squared Error over Different Test-Train Splits: Mean squared error is another measure for evaluating whether the linear model performs well. In other words, we performed cross validation, which is randomly splitting the dataset into K subsets, and each time one of the K subsets is the test set and the other K-1 subsets together is the train set. Here, we chose K=10 (out of popular choice). For each of the test-train splits of the avocado dataset, we first fit a linear model using all predictors except for date and year on the train set, which is our best-performing model. Then, we recorded the Mean Squared Error between all pairs of predicted Average Price using the model fit before, and the true test Average Price. Since there is randomness, a seed was also set. This is the output of all 10 MSEs for one of the seeds: 0.07032501, 0.06974108, 0.07517237, 0.07584184, 0.06971868, 0.07269931, 0.07369682, 0.07262151, 0.07003823, and 0.07207937. The average of this list of errors is 0.07219342. Compared to the mean (=1.406) and variance (=0.162) of the column Average Price, the MSE is fairly small, indicating that the best linear model explored in this section is relatively good.

Analysis #2: Trend Fitting and Forecasting



[Figure 1a: Historical & Forecasted Average Prices]



[Figure 1b: Historical & Forecasted Average Sales Volume]

Average monthly prices model: From the exploratory data analysis (EDA) we've implemented, we found that there was a seasonal pattern that cycles throughout the year with regards to the average prices of an avocado (see Figure 3a). The best method to create a model that describes these findings is the Holt-Winters Exponential Smoothing. To obtain our model, we trained it using the average prices by month from 2015 to 2017 and used it to forecast the average prices throughout 2018 since our dataset ends in March 2018. The resulting model is displayed above as Figure 1a. We can see that our original and training data match relatively nicely. We went a step further and found the Sum of Squared Error (SSE), which was around 0.14165.

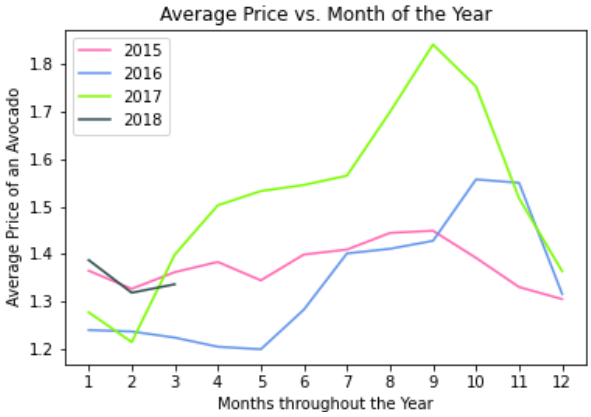
Average monthly sales volume model: Similarly, the seasonal pattern with average avocado sales volume (the number of avocados sold) throughout the year can be seen so Holt-Winters Exponential Smoothing was carried out again (see Figure 3b). This new model was trained with average monthly total volume data from 2015 to 2017. This model would help us predict the future sales volume through the entire year of 2018. The final is presented above as Figure 1b. Like the model with average prices, the original data and the training data match quite well. In fact, the original and training data match almost exactly for 2015. The SSE for this model was roughly 169343641744.39365.

Month	Year	Average Price
1	2018	1.31
2	2018	1.28
3	2018	1.35
4	2018	1.39
5	2018	1.38
6	2018	1.44
7	2018	1.49
8	2018	1.55
9	2018	1.60
10	2018	1.60
11	2018	1.50
12	2018	1.36

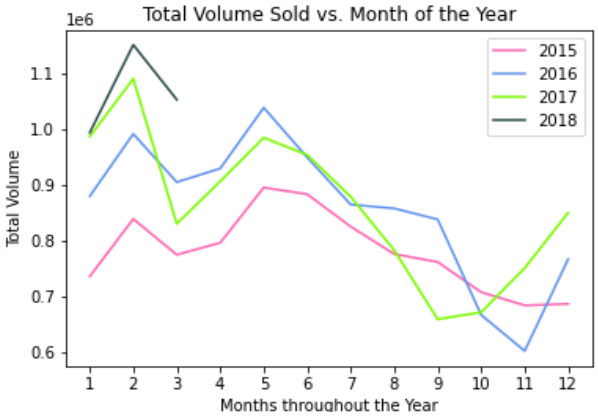
[Figure 2a: Forecasted Average Prices in 2018]

Month	Year	Sales Volume
1	2018	1.001925e+06
2	2018	1.074333e+06
3	2018	8.368274e+05
4	2018	9.110236e+05
5	2018	9.934212e+05
6	2018	9.523858e+05
7	2018	8.725710e+05
8	2018	7.986386e+05
9	2018	7.211060e+05
10	2018	7.238225e+05
11	2018	7.652757e+05
12	2018	8.157992e+05

[Figure 2b: Forecasted Sales Volume in 2018]



[Figure 3a: Average Prices of an Avocado vs. Month of Year]



[Figure 3b: Average Sales Volume vs. Month of Year]

Forecasting average monthly prices: From running the Holt-Winters Exponential Smoothing, we obtained a model that forecasts the average prices of an avocado throughout 2018, shown in Figure 2a. The minimum average price, $1.28 for an avocado, is expected to be in February 2018, while the maximum average price, $1.60 for an avocado, is expected to be in September and October 2018. These results align with the EDA we conducted prior where we examine the average prices of an avocado versus the month of the year, depicted by Figure 3a. For three of the four years of available data (years 2015, 2017, and 2018), February seems to be consistently the absolute minimum for average prices. Like Figure 3a suggests, the highest average prices of an avocado is usually around Fall months, September or October, for all four years. Furthermore, the SSE using this model is approximately 0.14165, which is extremely low. This means that the model we have is a good predictor of future average avocado prices.

We refer to “future” as in beyond the dataset we are analyzing, which only contains data that goes up to March 2018.

Forecasting average monthly sales volume: From the model on avocado sales volume in Figure 1b, we can predict that in 2018, we should expect to see the highest number of avocados sold in the early months of the year, namely January and February. In January 2018, we should anticipate approximately 1,002,000 avocados to be sold. In February 2018, we should anticipate the most sales volume, clocking in at approximately 1,074,000 avocados to be sold. The lowest number of avocados sold will occur in the fall months, September and October, with the least sales in September. In September 2018, we forecast that there will be around 721,100 avocados to be sold. In October 2018, we predict that there will be approximately 723,800 avocados to be sold. The SSE for this model, 169343641744.39365, is a lot higher than the previous model, but this may be due to the high variability in the ranges of values for total volume in our dataset. Figure 1b shows that our original and training data match fairly well.

It’s very interesting to see that there’s an inverse relationship between the number of avocados sold and the average price of an avocado in these key months. In February 2018, we anticipate a high sales volume and a low average price. Meanwhile, we believe that in September and October 2018, there will be low sales volumes and high average prices. It appears that as the price of avocados increases, the number of avocados sold decreases.

Conclusion: We can predict the monthly average prices and the sales volumes for the entirety of 2018, which is beyond the dataset we are examining. We should expect to see the most expensive avocados in September and October 2018, while the least expensive avocados in February 2018. The highest average price is $1.60 per avocado while the lowest average price is $1.28 per avocado. We have high confidence in these predictions given that they follow the trends established earlier by the EDA and the model has the low SSE value of 0.14165. We should expect to see the lowest number of avocados sold in September (only 721,000). The second lowest number of avocados sold will be in October, when 723,800 avocados will be sold. The most number of avocados sold will be in February 2018, with 1,074,000 to be sold. We are confident in these values as well despite the high SSE because the model in Figure 1b follows the trends from the EDA and the original and training data align with each other. There’s evidence that suggests an inverse relationship between the average prices and sales volumes. The remaining average prices and sales volume can be seen in Figures 2a and 2b.

Analysis #3: Logistic Regression

Logistic regression is a good choice when determining the effect of predictors on a binary categorical variable, such as the type (conventional or organic) of avocado sold variable in our dataset. In this section, we seek insight into the relationship between all predictors and avocado type.

Initial fit: We initially fit a logistic regression based on all variables excluding date and region. Excluding Date is due to the fact that logistic regression does not adequately analyze variables in datetime format. The model considers each datetime value as a separate level in a categorical variable. Additionally, we decided to leave out region in our regression as region is a categorical variable, and its effects on avocado

type would not be able to be quantified in logistic regression. It doesn't make sense to predict whether each location would buy conventional or organic using logistic regression.

Based on p-values, the predictors AveragePrice, SmallBags, LargeBags, and XLargeBags are the best predictors, although all except TotalBags have a p-value under 0.05. It is interesting that all the variables quantifying baggage size except TotalBags have a significant effect on the type of avocado. Perhaps this is because organic and conventional avocados are typically packed in a consistent bag size. However, it is unclear if this is due to the actual size of the different kinds of avocados, or the volume size of avocados shipped. This idea is expanded upon in a later regression.

Testing several combinations: We set up subsequent regressions comparing:

- type to AveragePrice
- type to SmallBags, LargeBags, and XLargeBags
- type to SmallBags, LargeBags, XLargeBags, TotalVolume
- type to TotalVolume

We believe that focusing on the pseudo R squared value for each regression run gives us the best indication of how accurate the prediction variables are in compensating for the observed variability in each generated model. Furthermore, we converted pseudo R squared to adjusted R squared, just to be sure that the increased accuracy of the model is not due to the increase in the number of predictors. We used a mathematical formula to convert pseudo R squared values to adjusted R squared values for each model generated: $1 - \frac{(1-R^2)(n-1)}{n-p-1}$, where n is the number of observations and p is the number of predictors. The adjusted R squared values are below:

1. (Initial fit) Model with type vs. all predictors except Date and region: 0.77115
2. Model with type vs. AveragePrice: 0.350993
3. Model with type vs. SmallBags, LargeBags, and XLargeBags: 0.45075
4. Model with type vs. SmallBags, LargeBags, XLargeBags, and TotalVolume: 0.687497
5. Model with type vs. TotalVolume: 0.492617

Actually, since there are 18,249 data points in the avocado dataset, the conversion from pseudo R squared value to adjusted R squared value does not change significantly (in fact, hardly changes a thousandth of a decimal place). We thought this could have been the case in our generated models, as the models with only 1 predictor have the smallest and third-smallest R squared values when compared to the other models with more predictors. That is why we computed adjusted R squared values of models to account for the different numbers of predictors.

Model comparison: Type vs. AveragePrice has a significantly smaller adjusted R squared value than type vs. TotalVolume (0.350993 < 0.492617). Therefore, it seems that the model solely including TotalVolume does a better job of explaining the variability exhibited in the data than the model generated with AveragePrice. It seems to logically follow that we can obtain a better idea for if an avocado is conventional vs. organic based on the volume of each type sold rather than the average price at which they are sold. Perhaps this is due to average price being regionally dependent, or Total Volume being directly related to supply and demand.

Difference of Additional TotalVolume: The model generated with predictors SmallBags, LargeBags, and XLargeBags has a much smaller adjusted R squared value than the model that includes predictors SmallBags, LargeBags, XLargeBags, and TotalVolume (0.45075 < 0.687497). The model with the additional TotalVolume appears to do a better job of explaining the variability exhibited in the data. Although the model with a higher adjusted R squared value has an additional predictor (TotalVolume), the difference between these models' adjusted R squared values seems to outweigh the difference of one predictor. This makes sense, as above we concluded that the volume of each type exported better indicates conventional vs. organic avocados. Including this in a model with baggage type should (and evidently does) increase the adjusted R squared value.

Dep. Variable:	type	**No. Observations:**	18249
Model:	Logit	**Df Residuals:**	18244
Method:	MLE	**Df Model:**	4
Date:	Mon, 17 May 2021	**Pseudo R-squ.:**	0.6876
Time:	16:28:44	**Log-Likelihood:**	-3951.7
converged:	True	**LL-Null:**	-12649.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.4833	0.038	64.621	0.000	2.408	2.559
TotalVolume	-3.435e-05	7.54e-07	-45.579	0.000	-3.58e-05	-3.29e-05
SmallBags	5.408e-05	1.51e-06	35.830	0.000	5.11e-05	5.7e-05
LargeBags	5.857e-05	1.6e-06	36.682	0.000	5.54e-05	6.17e-05
XLargeBags	-0.0367	0.002	-19.689	0.000	-0.040	-0.033

[Figure 4a: Logistic Regression on type using certain predictors]

Dep. Variable:	type	**No. Observations:**	18249
Model:	Logit	**Df Residuals:**	18238
Method:	MLE	**Df Model:**	10
Date:	Mon, 17 May 2021	**Pseudo R-squ.:**	0.7713
Time:	15:22:46	**Log-Likelihood:**	-2893.0
converged:	False	**LL-Null:**	-12649.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

[Figure 4b: Logistic Regression utilizing all predictors]

Comparing two models with largest R squared: Although the model which includes all the predictors except Date and region has the largest adjusted R squared value of 0.77115 (as depicted in figure 4b), we believe that a significant explanation for why it is so high is due to the far greater amount of predictors incorporated as compared to all of the other models. The model in figure 4a which includes SmallBags, LargeBags, XLargeBags, and TotalVolume with an adjusted R squared value of 0.687497 seems like the best generated model for explaining the variability in the data. We believe that baggage type (SmallBags, LargeBags, and XLargeBags) helps us predict if an avocado is conventional or organic based on the model's adjusted R squared value for these 3 indicators (0.45075). Coupling this with the predictor TotalVolume gives us a strong basis of predictors we can observe to indicate if an avocado is conventional or organic.

Bibliography

Avocados Dataset: https://www.kaggle.com/neuromusic/avocado-prices
Proof of proportionality of AIC and c_p:
https://rstudio-pubs-static.s3.amazonaws.com/324771_0bd880964f064c53a70e757d5ef39669.html
Relationship between adjusted R^2 and pseudo R^2:
https://towardsdatascience.com/the-complete-guide-to-r-squared-adjusted-r-squared-and-pseudo-r-squared-4136650fc06c
Python files: 3120 Final Project Python Files