Predicting Avocado Prices with Different Models

Rebecca Gordon[1]

[1] University of Oregon

Author Note

Correspondence concerning this article should be addressed to Rebecca Gordon.
E-mail: rebeccag@uoregon.edu

Predicting Avocado Prices with Different Models

The outcome variable

Across the entire United States, avocados are being sold everyday despite their fluctuating price. Our goal for this project was to create a machine learning model capable of accurately predicting a given state's avocado price over time across states based on. Avocado prices vary based on type (i.e., conventional vs. organic).

We will be answering the research question: What is the strongest predictor of avocado prices in the United States? Thus, our goal is to find the feature that most strongly predicts the price of avocados in the United States. The purpose of this project is to provide consumers, local grocers, and farmer's markets with a tool to predict avocado prices.

Description of the Data

Core features and descriptive statistics

We analyzed the avocado prices dataset retrieved from Kaggle and compiled by the Hass Avocado Board. The dataset consists of approximately 18,000 records over 4 years (2015 - 2018). The dataset contains information about avocado prices by type (organic or conventional), region purchased in the United States, total volume sold, and date sold.

```
## tibble [18,249 x 14] (S3: tbl_df/tbl/data.frame)
##  $ ...1        : num [1:18249] 0 1 2 3 4 5 6 7 8 9 ...
##  $ Date        : Date[1:18249], format: "2015-12-27" "2015-12-20" ...
##  $ AveragePrice: num [1:18249] 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1.02 1.07 ...
##  $ Total Volume: num [1:18249] 64237 54877 118220 78992 51040 ...
##  $ 4046        : num [1:18249] 1037 674 795 1132 941 ...
##  $ 4225        : num [1:18249] 54455 44639 109150 71976 43838 ...
##  $ 4770        : num [1:18249] 48.2 58.3 130.5 72.6 75.8 ...
```

```
##  $ Total Bags  : num [1:18249] 8697 9506 8145 5811 6184 ...
##  $ Small Bags  : num [1:18249] 8604 9408 8042 5677 5986 ...
##  $ Large Bags  : num [1:18249] 93.2 97.5 103.1 133.8 197.7 ...
##  $ XLarge Bags : num [1:18249] 0 0 0 0 0 0 0 0 0 0 ...
##  $ type        : chr [1:18249] "conventional" "conventional" "conventional" "conventi
##  $ year        : num [1:18249] 2015 2015 2015 2015 2015 ...
##  $ region      : chr [1:18249] "Albany" "Albany" "Albany" "Albany" ...
```

```
## # A tibble: 6 x 14
##      x1 date       average_~1 total~2 x4046  x4225 x4770 total~3 small~4 large~5
##   <dbl> <date>          <dbl>   <dbl> <dbl>  <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1     0 2015-12-27       1.33  64237. 1037. 5.45e4  48.2   8697.   8604.    93.2
## 2     1 2015-12-20       1.35  54877.  674. 4.46e4  58.3   9506.   9408.    97.5
## 3     2 2015-12-13       0.93 118220.  795. 1.09e5 130.    8145.   8042.   103.
## 4     3 2015-12-06       1.08  78992. 1132  7.20e4  72.6   5811.   5677.   134.
## 5     4 2015-11-29       1.28  51040.  941. 4.38e4  75.8   6184.   5986.   198.
## 6     5 2015-11-22       1.26  55980. 1184. 4.81e4  43.6   6684.   6556.   127.
## # ... with 4 more variables: x_large_bags <dbl>, type <chr>, year <dbl>,
## #   region <chr>, and abbreviated variable names 1: average_price,
## #   2: total_volume, 3: total_bags, 4: small_bags, 5: large_bags
```

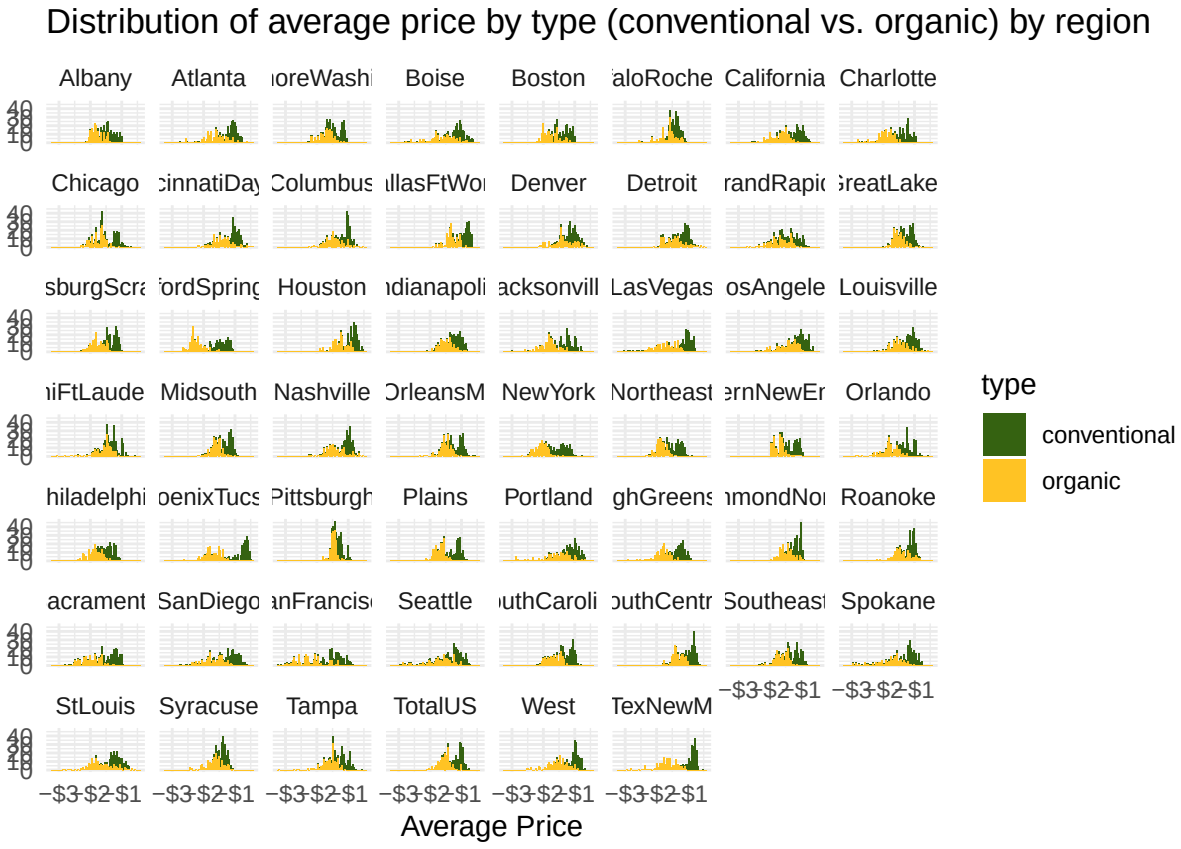
Frequencies and distribution of data

First, we subsetted the variables of interest from the dataset. From the histogram below, we can see that our outcome variable, average price, is normally distributed. Mean price across data was \$1.41 (SD = \$0.4). The highest average price for organic avocados was in San Francisco, CA in 2016 for \$3.25 and the lowest average price was in Cincinnati, OH in 2017 for \$0.44.

Table 1

Frequencies of the data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurt
date	1	18249	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	
type*	2	18249	1.50	0.50	1.00	1.50	0.00	1.00	2.00	1.00	0.00	-2
average_price	3	18249	1.41	0.40	1.37	1.38	0.42	0.44	3.25	2.81	0.58	0
total_volume	4	18249	850644.01	3453545.36	107376.76	232479.38	152652.16	84.56	62505646.52	62505561.96	9.01	92
region*	5	18249	27.50	15.58	27.00	27.50	19.27	1.00	54.00	53.00	0.00	-1



Missing data check

```
##            x1          date average_price  total_volume          x4046
##             0             0             0             0             0
##          x4225          x4770    total_bags    small_bags    large_bags
##             0             0             0             0             0
##  x_large_bags          type          year        region
```

```
##                0                0                0                0
```

No missingness was found for the variables in the dataset.

Description of the models

Three different modeling approaches will be used to predict avocado price from sale features, including: Linear Regression, Decision Trees, and Random Forest. Since the purpose of this project is to provide consumers, local grocers, and farmer's markets with a tool to predict avocado prices, we want to examine the predictive power of several features that contribute to avocado price. Thus, we first examined the effect of all predictors in a linear regression model to compare with the advanced models. Next, we added more complexity to the model by growing and pruning decision tree regression models to predict avocado price. Finally, we used a random forest to reduce the variance to get a more accurate prediction.

Model Fits

Preparation

The dataset is split into training and test set with the following code. The training set has 14,599 observations, and the test set has 4,650 observations. We examined which model has the optimal fit to predict avocado price (RMSE, MAE, and R-squared).

```
require(recipes)
loc <- sample(1:nrow(df), round(nrow(df) * 0.8))
df_train  <- df[loc, ]
df_test  <- df[-loc, ]


#dim(df_train)
#dim(df_test)
```

Model 1: Linear Regression Model with Cross Validation

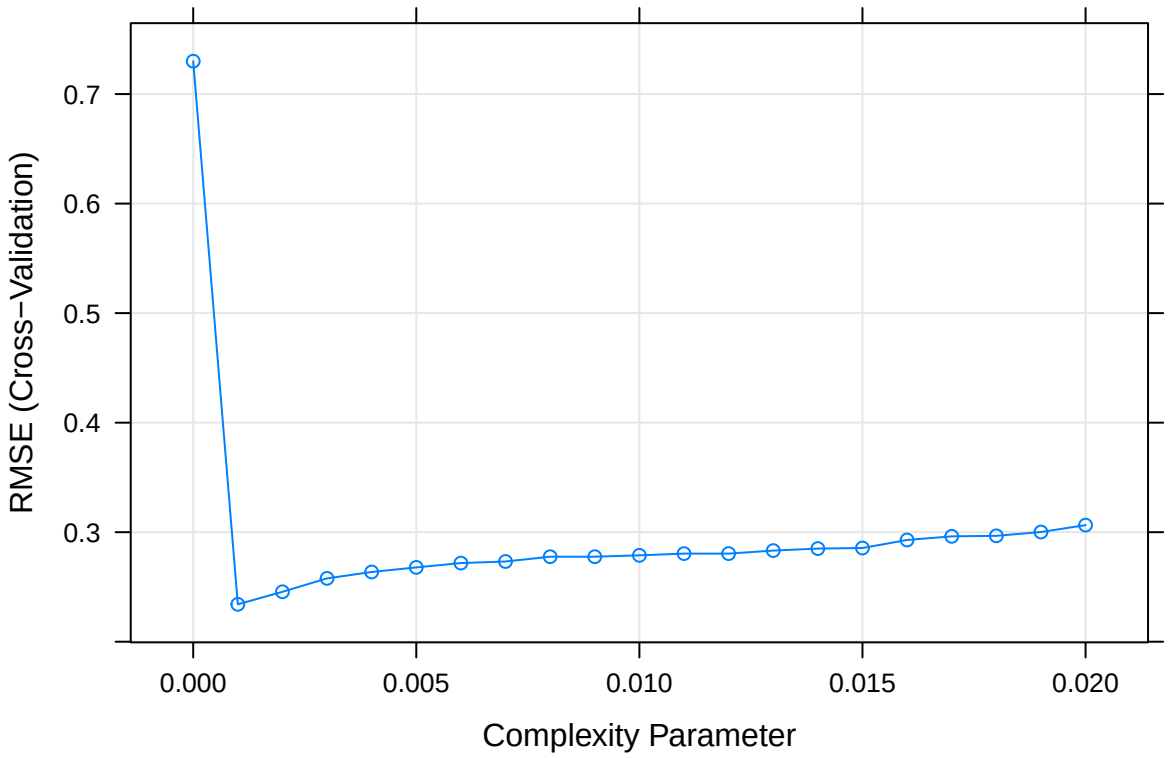
To obtain a general understanding of how our predictor variables fit our data, we first fitted a linear regression model without regularization. The equation generated by the linear model is then applied to predict outcome of new unseen data. Our criteria for evaluation of model fit will be the root mean square error (RMSE). We used 10-fold cross validation to determine the optimal value for the complexity penalization factor, alpha.

```
##   intercept      RMSE  Rsquared       MAE      RMSESD RsquaredSD       MAESD
## 1      TRUE 0.2633607 0.5689546 0.1988435 0.007748081   0.026833 0.005724263
```

Model 2: Decision Trees

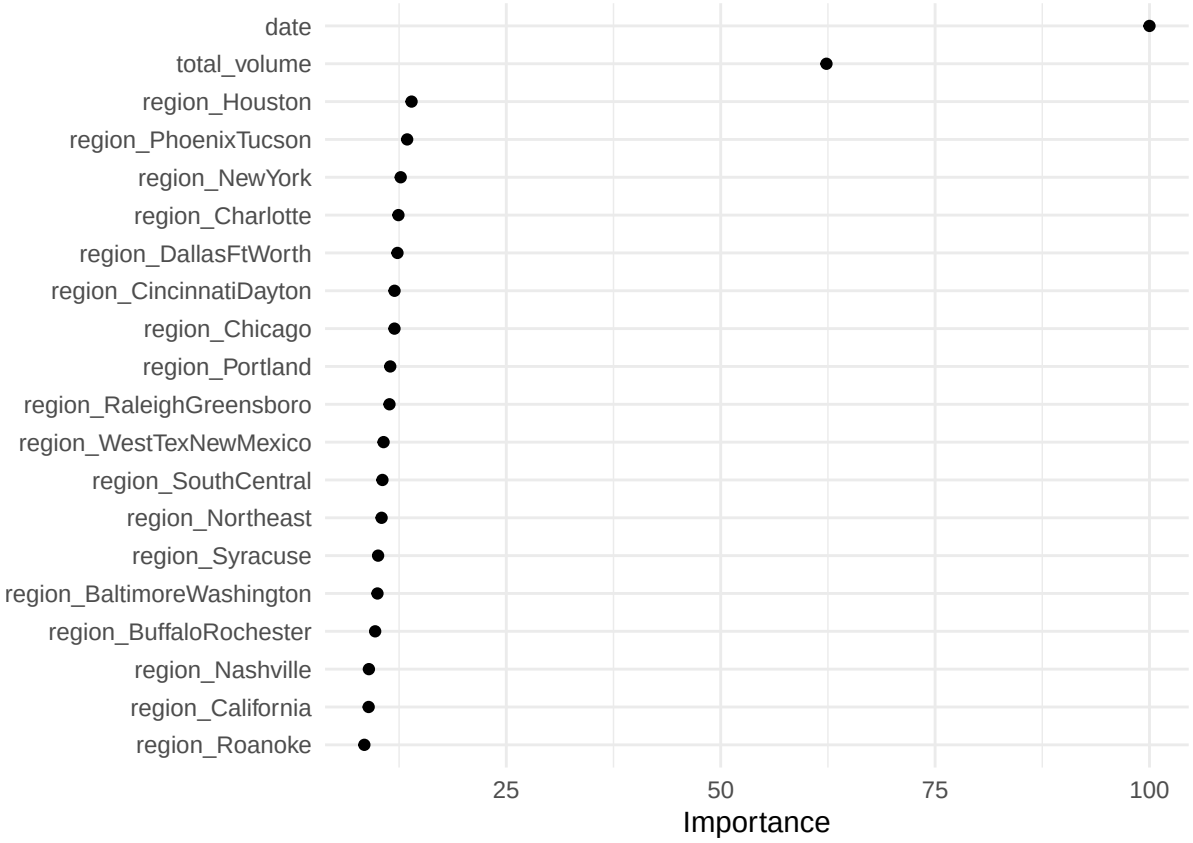
Next, we fitted decision tree regression model with cross validation to get a better estimate of the generalization error on new unseen data..

```
##   parameter   class                label
## 1        cp numeric Complexity Parameter
```



```
##       cp
## 2 0.001
```

We examined the complexity parameters for the model and found that date sold and total volume sold were the most important factors as predictors for average avocado price.



Model 3: Random Forest

Finally, we fitted a random forest regression model. According to the random forest regression, the top predictor of avocado prices is type (i.e. whether the avocado is organic or conventional). This result aligned with our expectations, as our preliminary data analyses depicted differences in distributions between organic and conventional avocado prices.

```
##   mtry splitrule min.node.size
## 1   25  variance             2
```

```
## Growing trees.. Progress: 95%. Estimated remaining time: 1 seconds.
## Growing trees.. Progress: 93%. Estimated remaining time: 2 seconds.
## Growing trees.. Progress: 100%. Estimated remaining time: 0 seconds.
## Growing trees.. Progress: 98%. Estimated remaining time: 0 seconds.
## Growing trees.. Progress: 83%. Estimated remaining time: 6 seconds.
```

```
## [1] 1.219530 1.189503 1.236844 1.242663 1.259437 1.222274
```

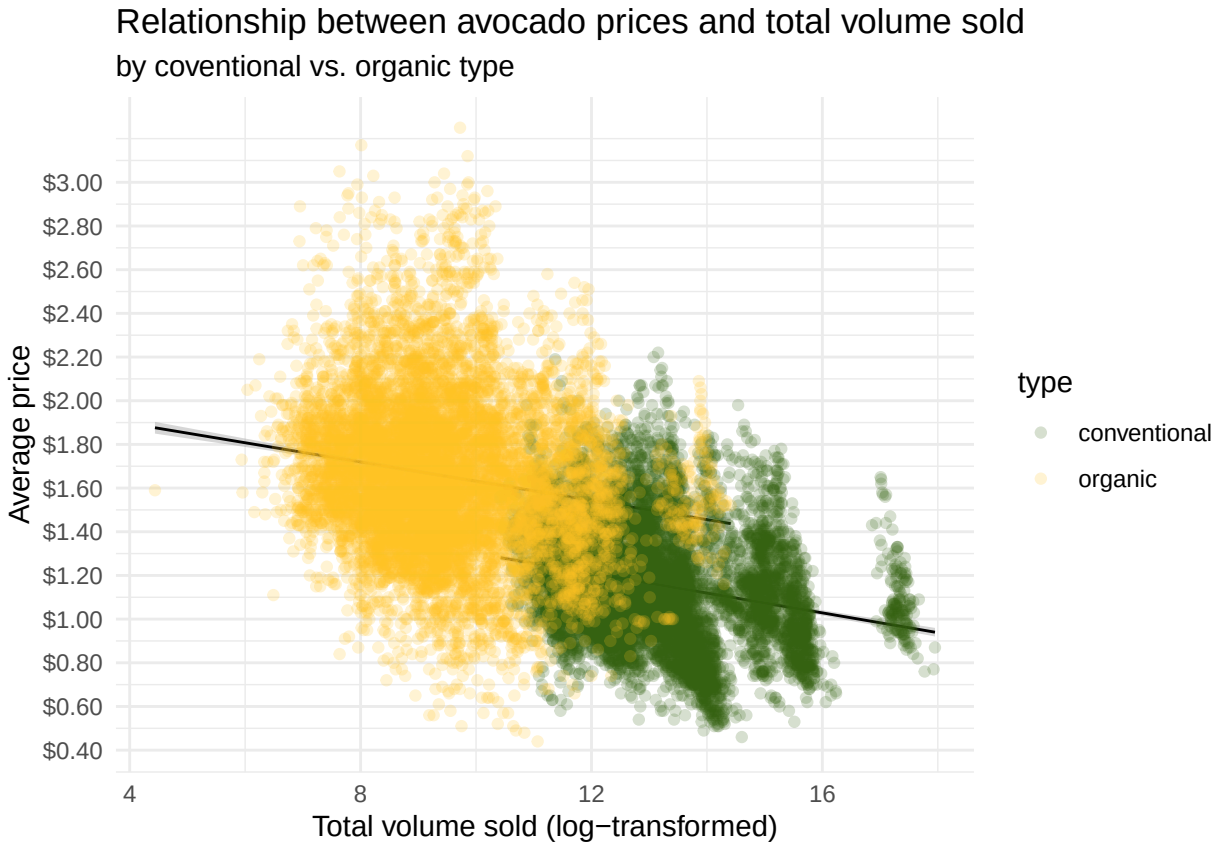
Comparing Models

Examining the predictions we can see that the random forest model outperformed the linear and decision tree models. This is because it has the highest R^2 and the least error. Thus, we can assume that random forest models can be trusted to predict avocado prices.

Model	Rsquare	RMSE	MAE
Linear Regression	0.5548706	0.2729584	0.2042162
Decision Trees	0.7022162	0.2233039	0.1709096
Random Forest	0.9136772	0.1221203	0.0857799

Data Visualization

Figure 1: Avocado Prices and total volume sold by type with regression lines. We examined our variables of interest visually with several plots. First we log-transformed the total volume sold to examine its relationship with average price. We can see that more conventional type avocados were sold at a lower price than organic avocados.


Relationship between avocado prices and total volume sold
by coventional vs. organic type
Average price
$3.00
$2.80
$2.60
$2.40
$2.20
$2.00
$1.80
$1.60
$1.40
$1.20
$1.00
$0.80
$0.60
$0.40
4
8
12
16
Total volume sold (log−transformed)
type
conventional
organic

126

Figure 2: Distribution of avocado prices by type.

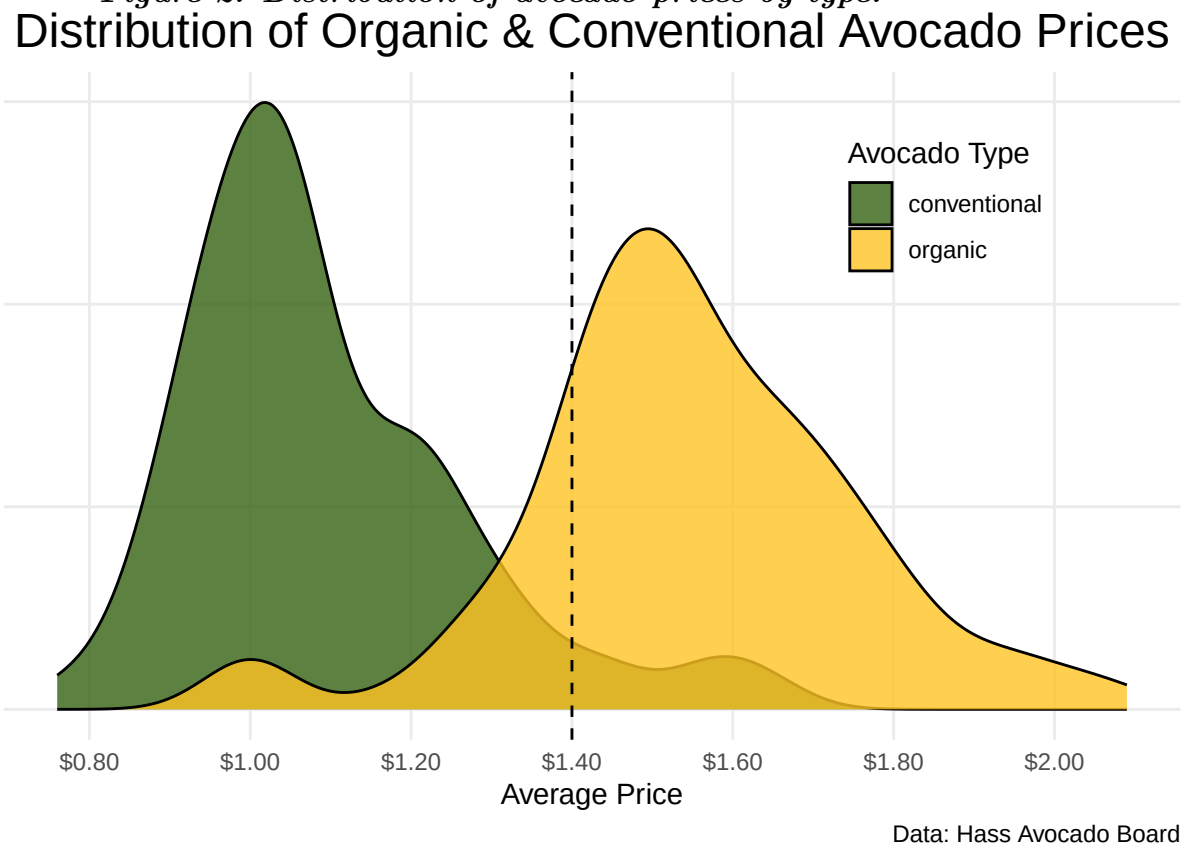


Figure 3: Regional avocado prices change over time.

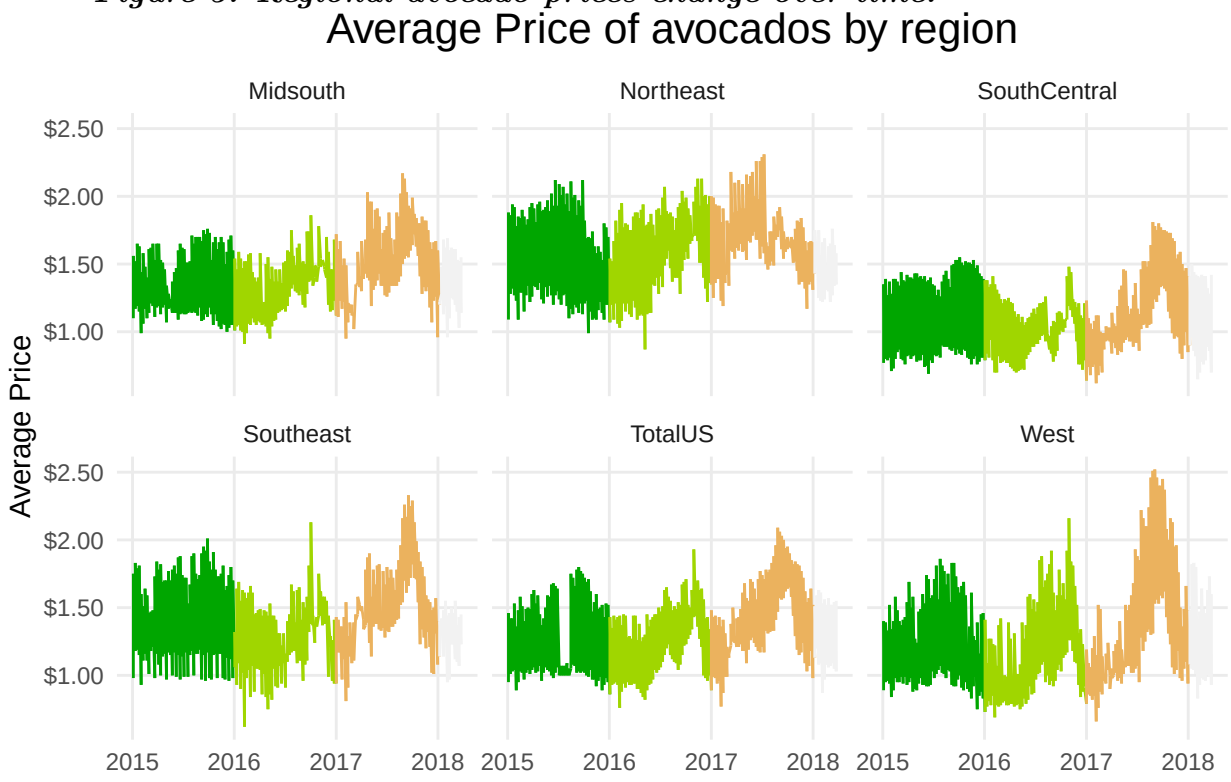
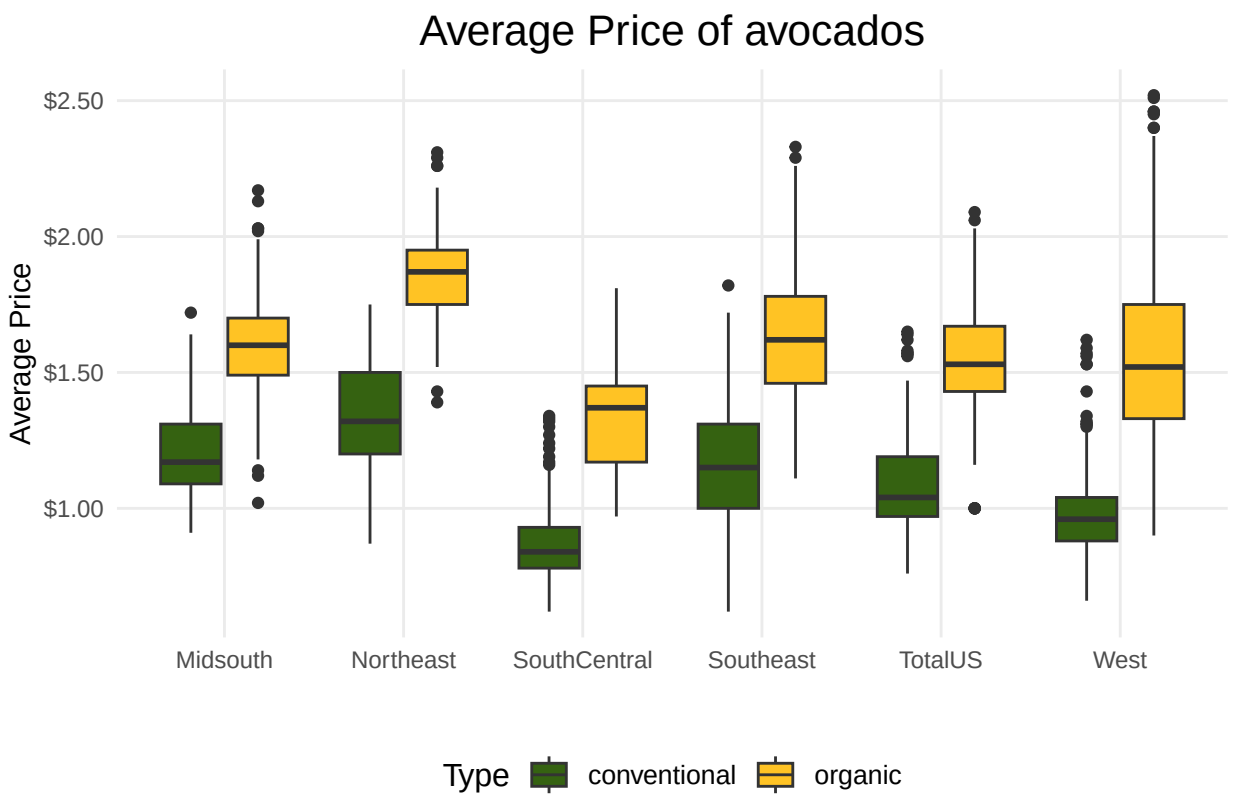


Figure 4: Boxplots of average price of avocados by region compared with total US.



Discussion

Conclusion

We optimized three models to predict average avocado prices and we found that the test scores for our predictive models were overall high. For the random forest regression model, the total variance explained was 91% and the decision tree model explained 69% of the variance.

We discovered some interesting findings from the models. The decision tree regression model predicted that date sold and total volume sold are the most important features for predicting avocado price.

The region where the avocado was sold was an important feature in the pricing of avocados. For instance, regions such as Baltimore/Washington and Houston were the third and fourth most important predictors of average avocado price.

References

Kiggins, J. (2018). *Avocado prices: Historical data on avocado prices and sales volume in multiple US markets*. Retrieved from https://www.kaggle.com/neuromusic/avocado-prices

Shahbandeh, M. (2019). *Average sales price of avocados in the u.s. 2012-2018*. Retrieved from https://www.statista.com/statistics/493487/average-sales-price-of-avocados-in-the-us/