M167R
Group 4

Hana Oh
Truc Nguyen
Jasmeen Kaur
Yuntian Yang
Qian Meng

Predict Average Price of Hass Avocado

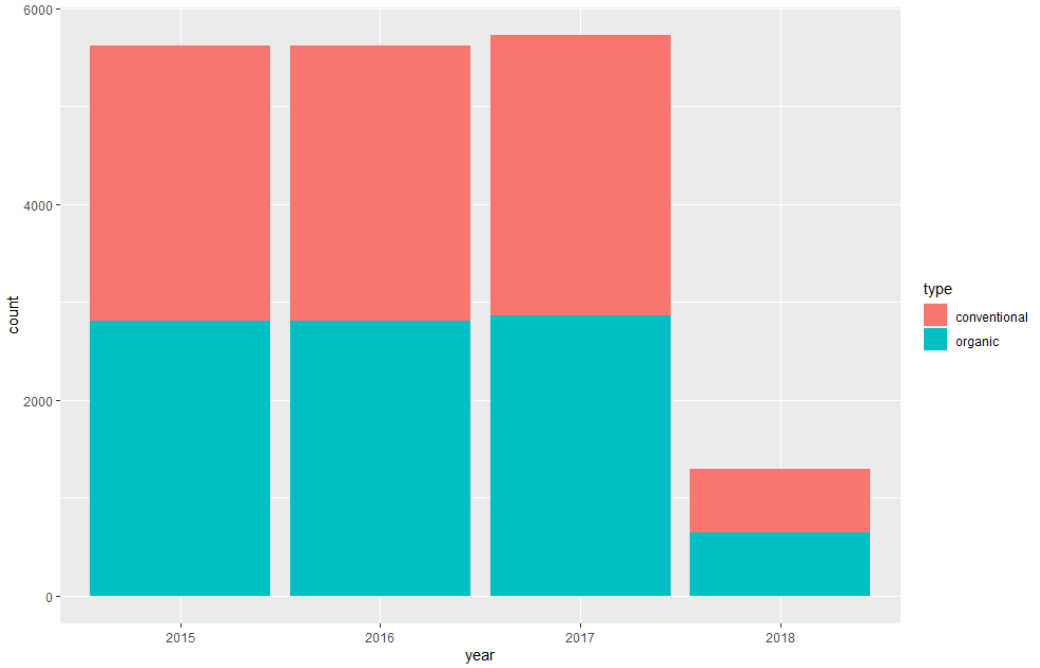
Introduction:

Our group is working on the Avocado dataset obtained from Hass Avocado Board website through Kaggle (Kiggins, 2018). The dataset contains information about the units of Hass avocados sold on various dates from 2015 to 2018 in several states, the quantities of three sizes of avocados sold and their average price. There are two types of avocados, organic and conventional. Our continuous predictor is date, and the categorical predictor is type. The response is average price. In this project, we build and test a multiple regression model to predict average price based on date and type.

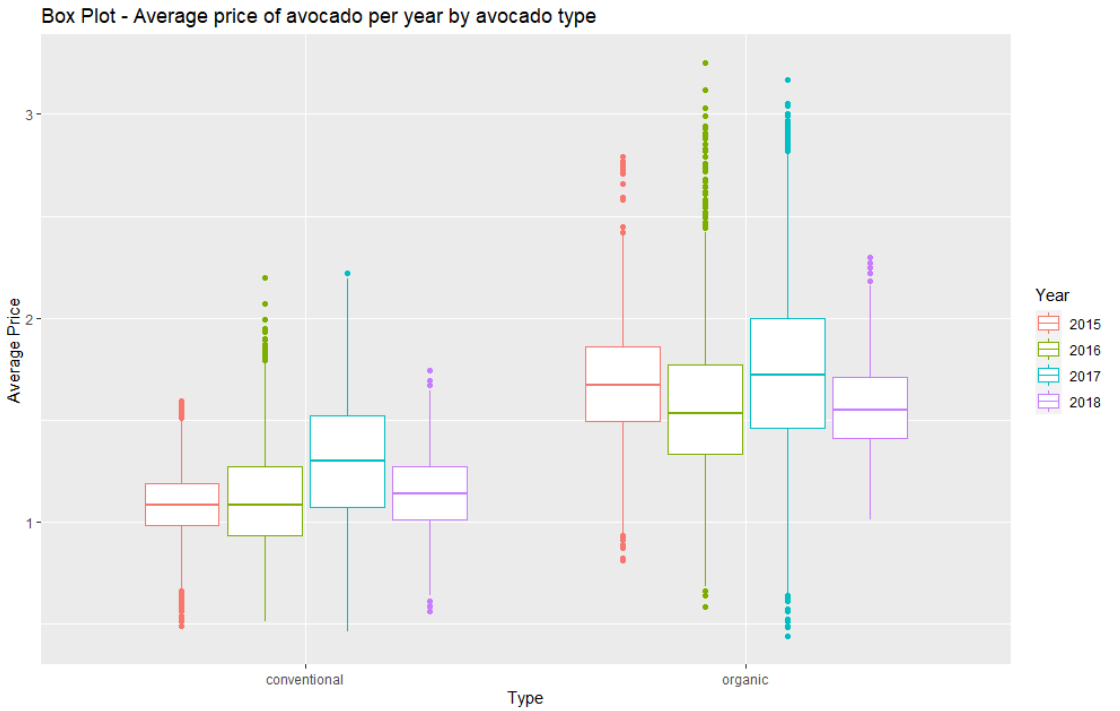
Variable name	Description	Range
AveragePrice	The average price of a single avocado in $	0.440 - 3.250
Date	The date of observation (each week)	2015/01/04 - 2018/03/25
type	The type of avocado 0: conventional 1: organic	

Even though date in our case is a discrete variable, we will see it as a continuous variable since it is an interval data. The model still needs some modifications but we learn that there is a relationship between date and type with the average price of avocado. From our research, the average price for organic avocados is consistent in this sample, while conventional avocados have some steep increases in average price.

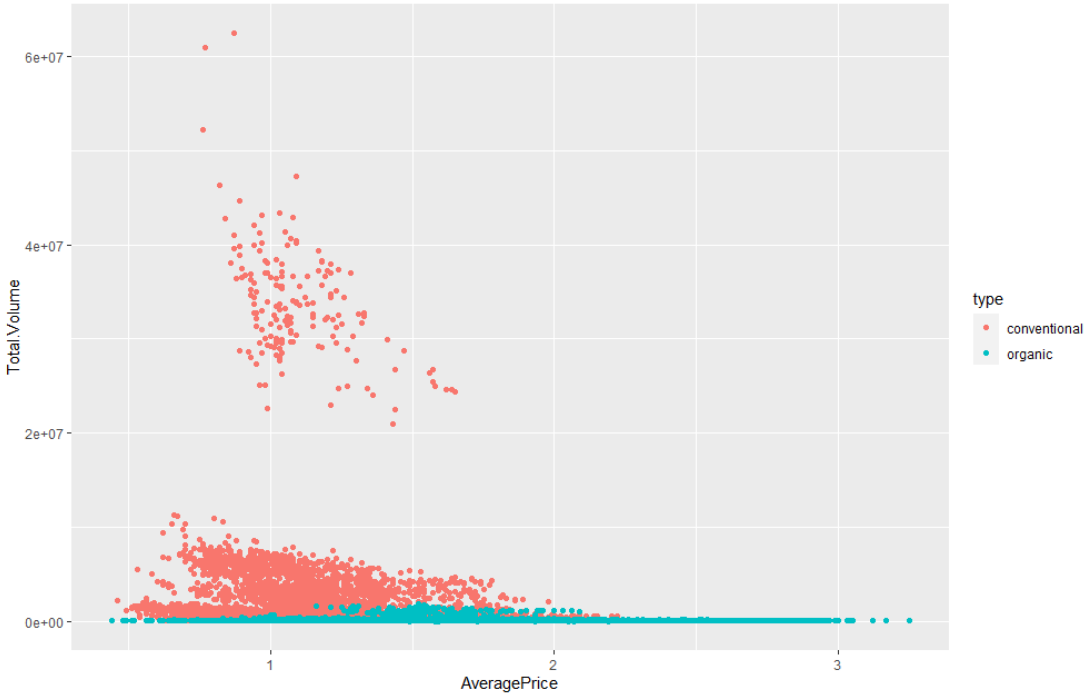
Scatterplot, boxplot, density plot, their outliers, and correlation heatmap: Yuntian Yang



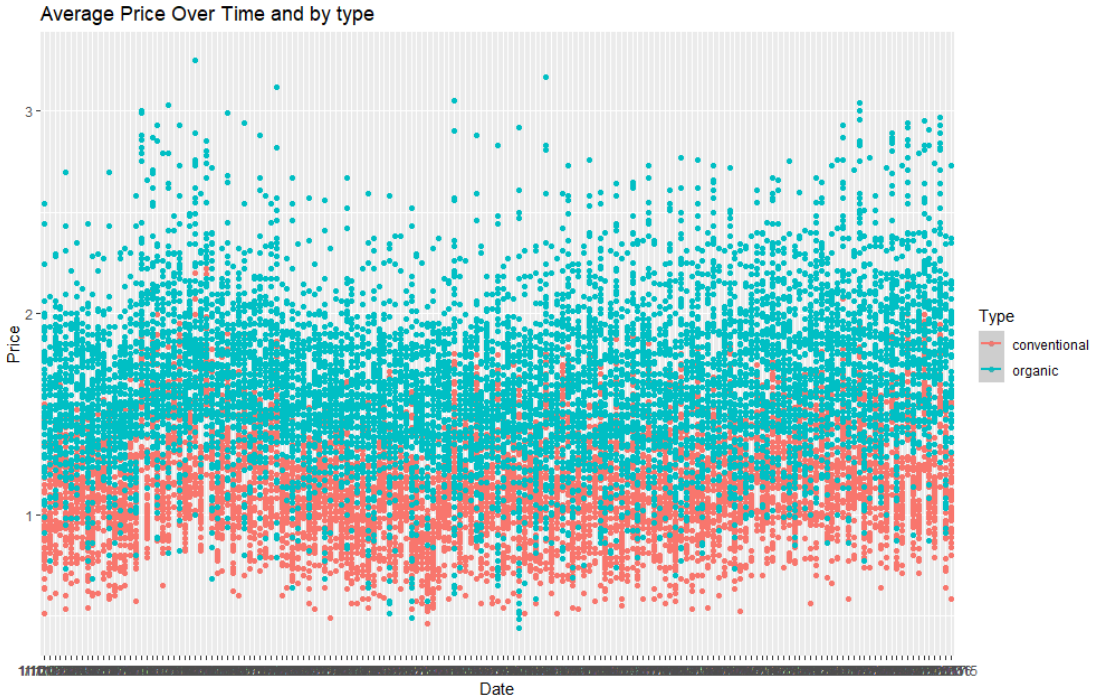
A barplot of the quantity of conventional and organic avocados from 2015 to 2018. The 2018 has a much shorter bar due to less data.



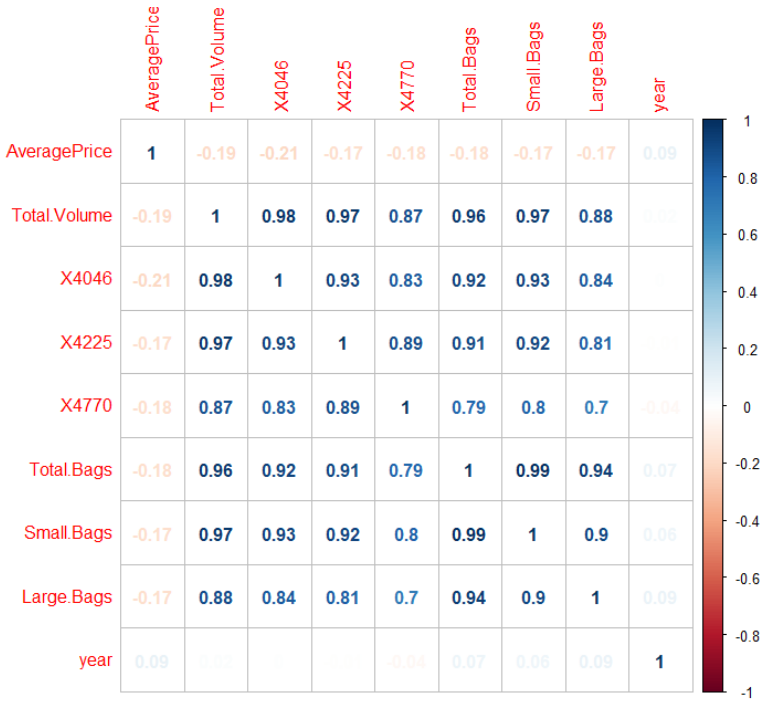
This boxplot compares the average price distribution between conventional and organic from 2015 to 2018. The points visible on the two ends of boxplots are outliers. We have 18,249 observations. The data is large enough and outliers are expected. Outliers are not excluded in the analysis.



A scatter plot to visualize the relationship between volume and average price by the two types of avocado.



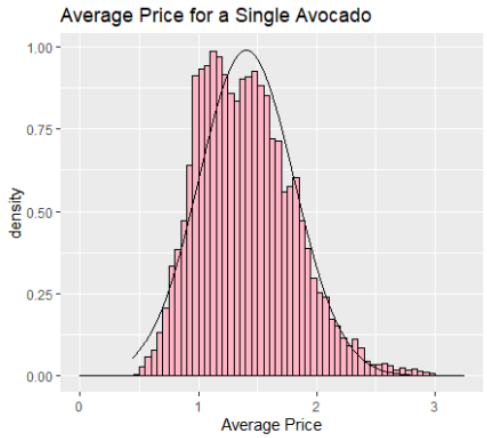
This scatter plot is to visualize the relationship between Average Price and Time by the two types of avocado.



The correlation heatmap shows some variables are highly correlated such as X4046 with Small Bags because Small Bag is a variable that depends on the value of X4046.

Histogram and building linear model: Truc

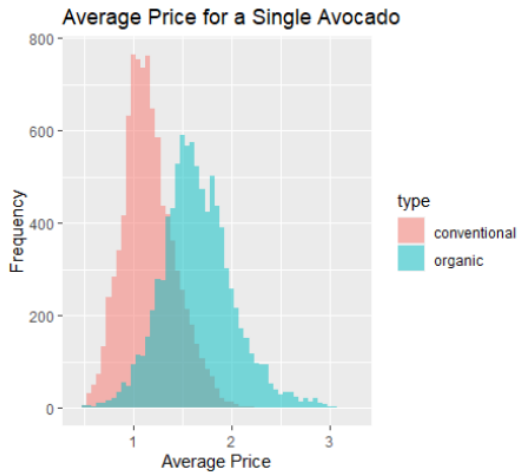
We investigate descriptive statistics to describe a sample. Below is the histogram of "Average Price for a Single Avocado".



First, analyze this histogram by creating descriptive statistics for this data:

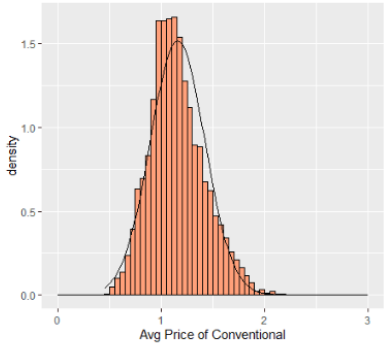
Statistic	Value
Mean	1.406
Range	1.100-1.660
Proportion>=75%	1.660
Standard Deviation	0.4026766
Median	1.370

These results indicate that the mean of average price for this sample is $1.406, with standard deviation of 0.402. We see that there is a broad peak in the histogram. It gives us a hint that this may be a multimodal distribution. We try dividing the sample by type: conventional and organic to see if type is the reason why we have the peak dividing into two parts in the histogram above.

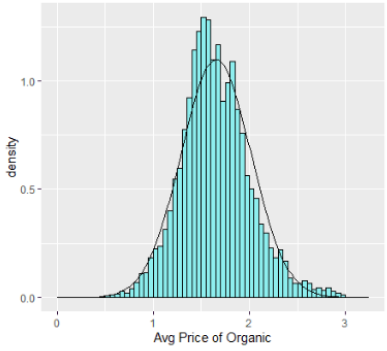


This histogram shows the average price for a single avocado by type. We can see that the histogram has two peaks, which means if we examine the mean by type, we will get a more precise estimate. Also, we expect the organic avocados to have a larger spread. This can be explained by the change in consumer's perception about organic avocado and health. Thus the price for organic avocado drops when demand shifts, so the distribution of its price is wider.

Below are two histograms that show the distribution of average price for conventional and organic avocados fitted under the normal curve.



Min. 1st Qu. Median Mean 3rd Qu. Max.
0.460 0.980 1.130 1.158 1.320 2.220
StdDev
0.2630406



Min. 1st Qu. Median Mean 3rd Qu. Max.
0.440 1.420 1.630 1.654 1.870 3.250
StdDev
0.3635016

The mean average price for each type ($1.158 & $1.654) is not the same as the mean of the sample ($1.406). So we want to test if type is a significant variable. Because organic and conventional avocado sales are independent, we use unpaired t-test. Since the sample size is large, we assume two groups of the sample are normally distributed.

We do a 2-sample two-tailed t-test to test whether the two groups means are equal.
Null hypothesis: Two means are equal.
Alternative hypothesis: Two means are not equal.

We then get p-value = 2.2×10^{-16} << significance level $\alpha = 0.05$, so we reject the null hypothesis. Thus, there is not enough evidence to conclude the mean of average price of conventional avocado is

equal to the mean of average price of organic avocado. That's why type is an important categorical variable in studies the average price of avocado.

From scatter plots above, we can see the relationship between average price of avocado and date. We will build a linear regression model and check if there is a significant relationship between date and type with the average price of avocado.

Consider:

Y = average price X_1 = date

Based on type variable, we create a new dummy variable that takes values as below:

X_2 = 0 if the avocado is conventional and 1 if the avocado is organic

Multiple regression model:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.056e+00	5.179e-03	203.87	<2e-16 ***
Date	1.201e-03	4.731e-05	25.39	<2e-16 ***
Typeorganic	4.960e-01	4.616e-03	107.44	<2e-16 ***

…

F-statistic: 6093 on 2 and 18246 DF, p-value: < 2.2e-16

Since in the F-test, p-value = 2.2 X 10^{-16} << significance level α = 0.05, there is enough evidence to conclude that our model with date and type has the better fit than the intercept-only model. The coefficients we get from R is the estimated intercept $\widehat{\beta_0}$= 1.055927, slope $\widehat{\beta_1}$= 0.001201, and slope $\widehat{\beta_2}$= 0.495966.

If the avocado is organic, we have the model:

$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}.(date) + \widehat{\beta_2}.1 + \epsilon \Rightarrow \hat{Y} = 1.055927 + 0.001201.(date) + 0.495966 + \epsilon$

If the avocado is conventional, we have the model:

$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}.(date) + \widehat{\beta_2}.0 + \epsilon \Rightarrow \hat{Y} = 1.055927 + 0.001201.(date) + \epsilon$

From the output, the average price for organic avocado is significantly associated with an average increase of *$0.495966*in average price compared to the conventional avocado.

We then check assumptions for these models.

<u>Model checking: P-value and t-value & r-squared and adjusted r-squared: Jasmeen</u>

The summary of our model returns numerous statistical values for the intercept, Date and Organic avocado type. The p-values are listed on the right-most column next to the stars. Since all of the p-values for each coefficient are well below the standard alpha level of 0.05, we can reject the null hypothesis that the predictor coefficients are zero. By rejecting the null hypothesis, we can conclude that there is a relationship between the independent variables and the dependent variable for our model. The three stars next to the p-values indicate how significant the variables are, with more stars indicating a higher significance.

```
> summary(fit)

Call:
lm(formula = AveragePrice ~ Date + type, data = avocado_2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24884 -0.20040 -0.01754  0.18628  1.58278

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.764e+00  1.151e-01  -15.32   <2e-16 ***
Date         1.716e-04  6.759e-06   25.39   <2e-16 ***
typeorganic  4.960e-01  4.616e-03  107.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3118 on 18246 degrees of freedom
Multiple R-squared:  0.4004,    Adjusted R-squared:  0.4004
```

The t-values are the calculated differences in units of standard error. Larger t-values indicate greater evidence against the null hypothesis. This means that there is enough evidence to imply that the coefficients are not equal to zero by pure chance alone. In our summary, we see that the t-values of Date and Organic-type are 25.39 and 107.44 respectively.

In addition to p- and t-values, the summary also returns Multiple R-squared and adjusted R-squared values. Multiple R-squared is a measure of how close our data is to the fitted regression line from our model (also known as the coefficient of determination). It ranges from 0 to 1 and explains how closely the variation in y can be explained by the x-variables. In our case, the Multiple R-squared value is 0.4004. Meaning, that 40.04% of the variation in Price can be explained by the Type of avocado and the Date it is purchased on.

Multiple R-squared does help to measure variation in the data however it is not an ideal form of representation. As more predictors are added, the R-squared value will only increase and imply that we are getting a better fit. However, this is only due to the increase in predictor variables and doesn’t help to determine if we are getting an actual better fit. These concerns are addressed by the adjusted R-squared value, as it only increases if the new predictor variables enhance our model. A large difference between the R-squared and adjusted R-squared values would indicate that we may have overfitted our data to the model. However, the values for both of these measures are equal. All these measures assure us that the model we have built is efficient for our dataset and any initial assumptions have been met.

Organic

```
lm(formula = AveragePrice ~ Date, data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23701 -0.23365 -0.02766  0.21695  1.58500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.386e-02  2.123e-01   0.159    0.873
Date        9.537e-05  1.247e-05   7.650 2.27e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3635 on 7296 degrees of freedom
Multiple R-squared:  0.007957,  Adjusted R-squared:  0.00782
F-statistic: 58.52 on 1 and 7296 DF,  p-value: 2.271e-14
```

Conventional

```
lm(formula = AveragePrice ~ Date, data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74730 -0.17289 -0.00592  0.15840  1.02378

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.211e+00  1.457e-01  -22.05   <2e-16 ***
Date         2.565e-04  8.553e-06   29.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2484 on 7298 degrees of freedom
Multiple R-squared:  0.1097,    Adjusted R-squared:  0.1096
F-statistic: 899.4 on 1 and 7298 DF,  p-value: < 2.2e-16
```

Above we have side by side comparisons of the summaries of the two models presented by Organic and Conventional avocado types. The p-values of the variable Date for both of these types is significantly smaller than 0.05. So, we can conclude that there is a correlation between Date and Price for both types of avocados.

We can also observe the p-values of the F-test in the last row. Since they’re both significantly less than 0.05, we can conclude that our regression models fit the data better than a model with no independent variables.

Model checking: Standard Error and F-Statistic, AIC & BIC:Qian

Then we go on to check the other 2 measures of goodness of model fit.

Standard Error （SE） is the standard deviation of the sample mean estimate of a population mean. Smaller SE indicates more approximating the sample mean is to the population mean. In our Organic (fit 1) and Conventional(fit 2) summary, we see that the SE of Date is 7.77e-05 and 5.32e-05 respectively. This means that the sample mean approaches the population mean.

```
> summary(fit1)

Call:
lm(formula = AveragePrice ~ Date, data = avocadoO)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23218 -0.23409 -0.02702  0.21907  1.58910

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.601e+00  7.618e-03 210.123  < 2e-16 ***
Date        6.268e-04  7.773e-05   8.064 8.33e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3622 on 9121 degrees of freedom
Multiple R-squared:  0.007079,  Adjusted R-squared:  0.00697
F-statistic: 65.02 on 1 and 9121 DF,  p-value: 8.334e-16
```

```
> summary(fit2)

Call:
lm(formula = AveragePrice ~ Date, data = avocadoC)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74720 -0.17122 -0.00622  0.15831  1.02243

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.007e+00  5.223e-03  192.81   <2e-16 ***
Date        1.776e-03  5.329e-05   33.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2484 on 9124 degrees of freedom
Multiple R-squared:  0.1085,    Adjusted R-squared:  0.1084
F-statistic:  1110 on 1 and 9124 DF,  p-value: < 2.2e-16
```

F-Statistic (F-test) indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables.

The F-test for overall significance has the following two hypotheses:

* The null hypothesis states that the model with no independent variables fits the data as well as your model.

* The alternative hypothesis says that your model fits the data better than the intercept-only model.

```
> summary(fit)

Call:
lm(formula = AveragePrice ~ Date + type, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24884 -0.20040 -0.01754  0.18628  1.58278

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.056e+00  5.179e-03  203.87   <2e-16 ***
Date        1.201e-03  4.731e-05   25.39   <2e-16 ***
typeorganic 4.960e-01  4.616e-03  107.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3118 on 18246 degrees of freedom
Multiple R-squared:  0.4004,    Adjusted R-squared:  0.4004
F-statistic:  6093 on 2 and 18246 DF,  p-value: < 2.2e-16
```

Here, the P-value of F-test is <0.05, means our model fits the data better with independent variables.

Next, we calculate AIC and BIC to see whether our model includes overfitting parameters. As listed below, the value of AIC and BIC of Total data(fit), Organic (fit 1) and Conventional(fit 2) are not significantly small. However, we use the only 1 continuous parameter(Date) and 1 dummy (type) in the model, the overfitting parameters seems not a big problem.

```
> AIC(fit)  # AIC(Akaike’s information criterion) => 9260.3
[1] 9260.3
> AIC(fit1)  # AIC(Akaike’s information criterion) => 9260.3
[1] 7365.713
> AIC(fit2)  # AIC(Akaike’s information criterion) => 9260.3
[1] 481.0699
> BIC(fit1) # BIC(Bayesian information criterion) =>9291.548
[1] 7387.068
> BIC(fit2) # BIC(Bayesian information criterion) =>9291.548
[1] 502.4265
> BIC(fit) # BIC(Bayesian information criterion) =>9291.548
[1] 9291.548
```

Predicting Linear Models

Starting with the Organic set.

In order to see how our prediction model will perform with new data, we split the dataset to 80:20. 80% of the dataset is used to train the model and another 20% is used as a test model.

AIC (Aikake Information Criterion)

Upon making assumptions that we are using the same data between the models, measuring the same outcome variables between models and having a sample of infinite size, it is used to observe the fit of the data based on its information loss. The lower the number is, the better fit data is which means it is more likely to minimize the information loss.

```
> AIC(lmMod) #akaike information criterion.
[1] 5985
```

We could use AIC to pick better model, but since we only have one we are just observing the number between the two types.

Correlation Accuracy

```
> correlation_accuracy
            actuals predicteds
actuals       1.000      0.119
predicteds    0.119      1.000
```

Our correlation accuracy between predicted values and actual values is approximately 12%. This means that the movement behavior of actual data is not very similar to the behavior of the predicted data.

Min Max Accuracy

MinMax Accuracy measures how accurate our predicted datas are to the actual data. It will compute accuracy rate of each row then take a mean of that by comparing min and max.

$$MinMaxAccuracy = mean\left(\frac{min\,(actuals, predicteds)}{max\,(actuals, predicteds)}\right)$$

$$MeanAbsolutePercentageError\,(MAPE) = mean\left(\frac{abs\,(predicteds - actuals)}{actuals}\right)$$

```
> #calculate min max accuracy and MAPE
> min_max_accuracy = mean(apply(actuals_preds, 1, min)
+                         /apply(actuals_preds,1,max))
> min_max_accuracy
[1] 0.848
```

Our MinMax Accuracy is approximately 84.8%, with 100% being perfectly accurate, our data prediction values are close to that of the actual dataset.

MAPE (Mean Absolute Percentage Error)

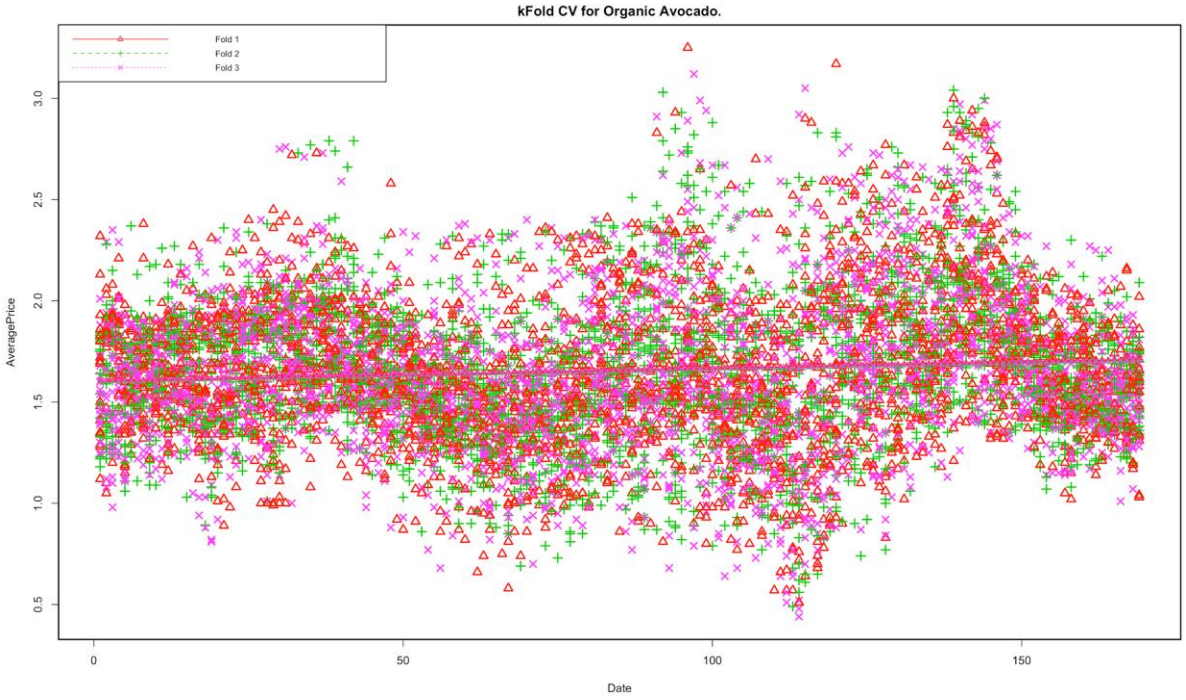
The **mean absolute percentage error** (MAPE) represents our model's prediction accuracy. The percentage represents that of error in the model.

```
> mape = mean(abs((actuals_preds$predicteds - actuals_preds$actuals))
+             /actuals_preds$actuals)
> mape
[1] 0.188
```

Our MAPE is approximately 18.8%, which shows that our model will accurately predict the average price close to 18.8% of the time in arbitrarily large trials.

kFold Cross Validation

This will test the stability of our prediction model on new data. In our project, we split the actual data into 3 folds (k = 3). This will use two other folds for training and one fold for testing in each iteration, and will proceed through each iteration by changing training folds and a testing fold.



In the chart above, we can see that the dashed lines in the center are nearly parallel with each other. This suggests that our model's predictions do not vary that much on a particular sample, and our colored symbols are generally spread in similar fashion.

Conventional set.

Aikake

```
> AIC(lmMod) #akaike information criterion.
[1] 395
```

Conventional avocados linear model has significantly smaller AIC value than of the organic. Thus, it is less likely to lose information.

Correlation Accuracy

```
> correlation_accuracy
            actuals predicteds
actuals       1.000      0.352
predicteds    0.352      1.000
```

We have 35.2% correlation accuracy between actual data and predicted data on conventional type of avocados average price based on the dates. This means that the trend of actual data and predicted data are not very close with each other.

MinMax Accuracy

```
> min_max_accuracy
[1] 0.848
```

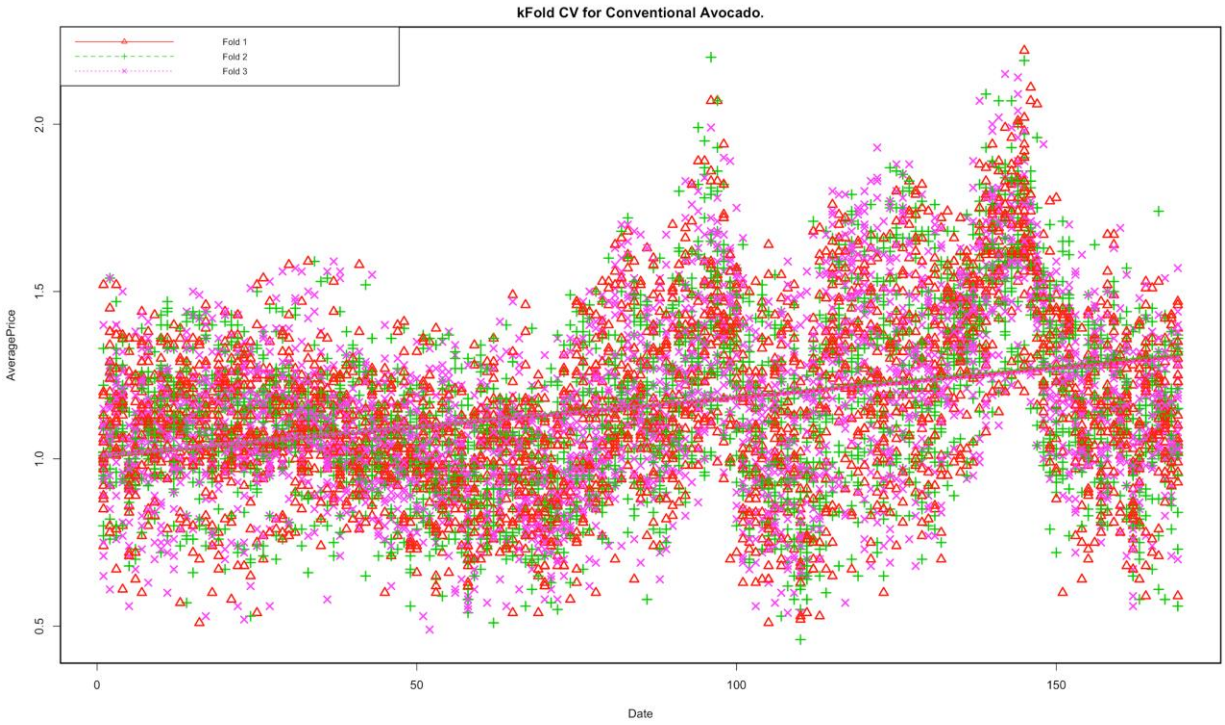
Our MinMax Accuracy is 84.8% for conventional avocados, thus our predicted values and actual dataset are close to each other in each date. This MinMax accuracy for conventional is approximately the same as that of organic avocados.

MAPE

```
> mape
[1] 0.184
```

For conventional avocados, MAPE is about 18.4%, which is slightly lower than organic avocados MAPE. In general, our prediction model for conventional avocado will have 18.4% of error in a large set of repeated samples.

kFold Cross Validation



In kFold CV for conventional avocados, the lines are more highly sloped than the organic avocados. The predicted values show spike on date 90 and 140, which would contribute to steeper increase in average price for conventional avocados. The dashed lines are all about parallel, showing that the predictions of sample sets did not vary from each other. Thus, both of our prediction models show consistency in its predictional.

Conclusion:

The model still needs some modifications and better analysis to predict the price of Hass avocado, such as time series analysis. But in the limit of what we have learnt, we discover the relationship between date and type with the average price of avocado by using the multiple regression model.

Citations:

Kiggins, Justin. “Avocado Prices.” Kaggle, 6 June 2018, www.kaggle.com/neuromusic/avocado-prices.