# Data Mining Report – Haas Avocado’s

John Griffin 16325223

Killian Byrne: 16781985

Submission Date 17/12/2019

Table of Contents

[Data Mining Report – Haas Avocado’s 1](#_Toc27582132)

[Section 1: Idea and dataset description 2](#_Toc27582133)

[1.1: Introduction 2](#_Toc27582134)

[1.2: The dataset 2](#_Toc27582135)

[Section 2: Data Preparation 3](#_Toc27582136)

[2.1: Learning about our dataset using *Python* 3](#_Toc27582137)

[2.2: Measuring the Dispersion of our Data (Our workshop) 4](#_Toc27582138)

[**2.3:** Feature Engineering 5](#_Toc27582139)

[**2.4:** Multicollinearity 6](#_Toc27582140)

[**2.5:** Data Standardisation 6](#_Toc27582141)

[Section 3: Algorithm description 7](#_Toc27582142)

[**3.1:** Multiple Linear Regression 7](#_Toc27582143)

[3.2: Ordinary Least Squares algorithm 8](#_Toc27582144)

[Section 4: Implementation, Results, Analysis & What we Learned 9](#_Toc27582145)

[4.1: Splitting our dataset into Training and Test data sets 9](#_Toc27582146)

[4.2: Feature selection & Linear Regression Calculation 9](#_Toc27582147)

[**4.3** What we learned whilst carrying out these experiments 13](#_Toc27582148)

[4.3.2 Conclusions 14](#_Toc27582149)

[Section 5: References 15](#_Toc27582150)

## Section 1: Idea and dataset description

### 1.1: Introduction

The dataset we have chosen for our CA4001 Data Warehousing and Data Mining is a dataset on the historical data on avocado prices and sales volume in US markets. The dataset was acquired from Kaggle and was created by Justin Kiggins. The data itself originated from the Hass Avocado Board website and was downloaded in May of 2018.

In terms of technologies used for each phase of our data analysis, we mainly utilised Microsoft Excel/LibreOffice Calc for some basic manipulation and calculation and used Python, with the libraries *Numpy* and *Pandas* for more advanced methods. We also utilised *seaborn* to help us with our statistical data visualization and *scikit-learn* for assistance with feature selection.

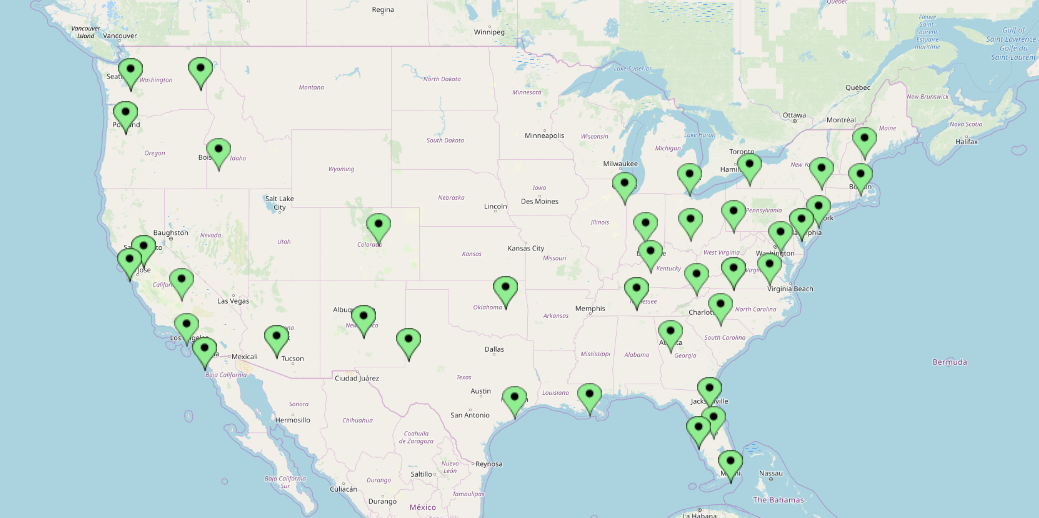
### 1.2: The dataset

#### **1.2.1** How was the data collected?

The dataset contains 18249 unique entries, spanning from year 2015 to May of 2018, which we believe is enough scope to make a prediction from. The dataset itself consists of ten attributes, which gives us a very large scope for our analysis and provides many analytic opportunities. The attributes are follows:

1. *Date* - The date of the observation, in the format [DD/MM/YYYY].
2. *Average Price* - The average price for the avocado, in US dollars.
3. *Total Volume*- The total number of avocados sold for the observation.
4. *4046* - Total number of avocados with PLU\* 4046 sold.
5. *4225* -Total number of avocados with PLU\* 4225 sold.
6. *4770* -Total number of avocados with PLU\* 4770 sold.
7. *Total Bags* - The total number of bags in the observation.
8. *Small Bags* - The total number of small bags in the observation.
9. *Large Bags* - The total number of large bags in the observation.
10. *Xlarge Bags* - The total number of xlarge bags in the observation.
11. *Type* - The type of Hass avocado, possible values are *conventional* or *organic*
12. *Year* - The year of the particular observation.
13. *Region* - The city or region of that particular observation.

\*PLU referring to *Price Look Up* Code. Here this helps the supermarket distinguish the size of the avocado, which is an interesting feature in itself: (Source: https://www.californiaavocado.com/retail/avocado-plus)

****

As we can see, the dataset is extremely well dispersed location wise throughout the United States of America. This provided us with the confidence that our results would in fact be a fair representation of avocado prices *throughout* America.

## Section 2: Data Preparation

### 2.1: Learning about our dataset using *Python*

*Brackets here refer to code snippets below.*

* A method that gives us more in depth information about our dataset is by utilizing the excellent *info()* method in Python’s Pandas library(1).
* We can see that float, int and object are the datatypes of our attributes.
* We see all our attributes have exactly 1849 entries, so this would indicate to us that there are no missing values. We employ the pandas *shape* to check what our data dimensionally is. The dimensionality should reflect the table in code snippet (2).
* These are identical. We can conclude there are no *missing* values or *blank entries* in our dataset (good indication of the quality of our dataset.)
* We then use the *describe()* method to give us the statistical characteristics of each numerical feature. These include all our attributes possessing the *Int* and *Float* type. The results of this analysis are as follows: (3)
* We now acquire the mean (for central tendency) and other attributes we are interested in analysing using this are our *object* attributes. This is acquired using the code snippet (4) and the results are shown in (5)
* We can see that there are 18249 entries into both *type* and *region.* Type, as expected contains two unique entries, namely *conventional* and *organic*.
* We gather also that there are 54 unique locations in our dataset.
* The conventional avocado is the most frequent type of avocado in our dataset,
* Jacksonville appears to be the most frequent region in our dataset. As it is well known that cities like New York and Los Angeles have considerably larger populations (CityMayors, 2019). We decided the best way to analyse this was to utilize *value\_counts()* in Pandas(6) and the results of this are found in (7).
* We see the reason for our previous output. Jacksonville was the most frequent because it was simply the first to appear when searching for a region.
* We now know that there are even representations for each region, and we could conclude that the extraction was carried out on a regular and consistent basis.
* We now convert the date object into a pandas readable datetime to make our analysis easier, and change the first column to a column called observation number.
* These values started as in (8) and after applying a pythonic solution, (9) our datetime was converted.
* To solve the problem of our observation number being “Unknown number” we had to employ a non-pythonic solution. Our found solution involved editing the .CSV file using LibreOffice Calc (10).

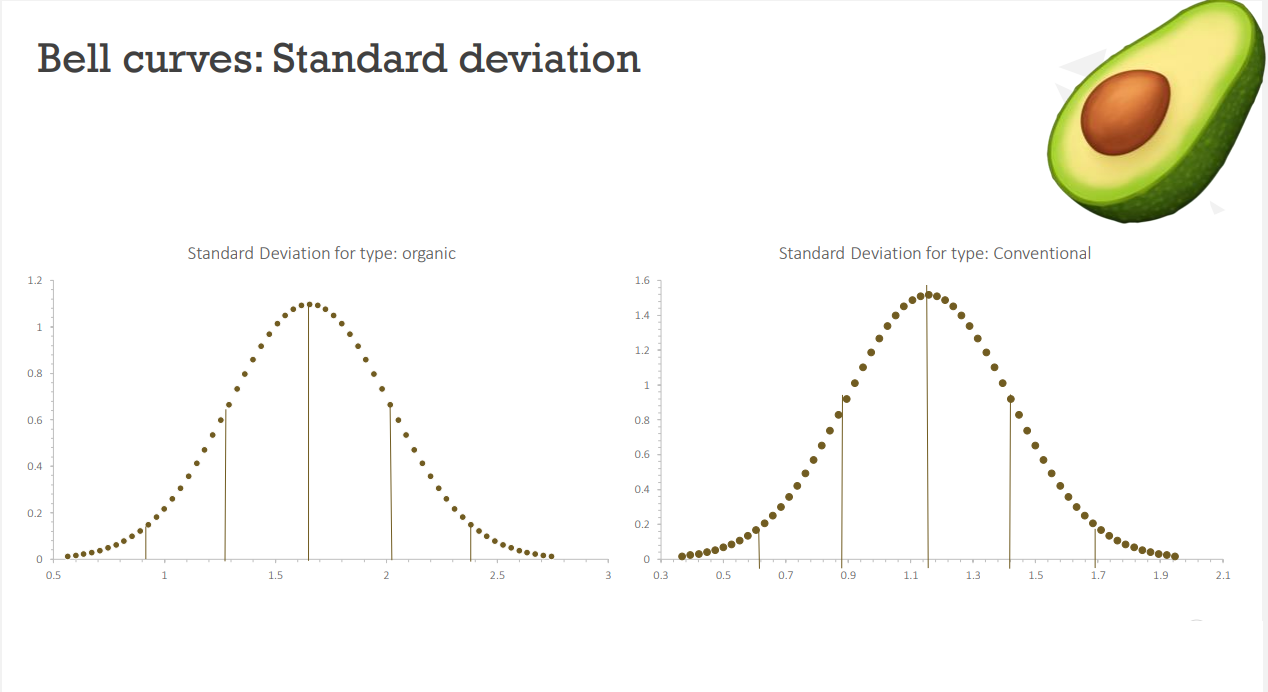
### 2.2: Measuring the Dispersion of our Data (Our workshop)

* We began the process of measuring the dispersion of our data by calculating the mean Avocado prices for organic and conventional avocados, for all the years included in our dataset. This was from the year 2015 to the year 2018.
* To measure the dispersion of, we carried out the following calculations on each dataset: the range, interquartile range, the five-number summary, standard deviation and variance, and the coefficient of variation.
* The five-number summary (Min, Q1, Med, Q3, Max) and the standard deviation (std dev), variance and coefficient of variation (CV) for the two datasets are as follows:

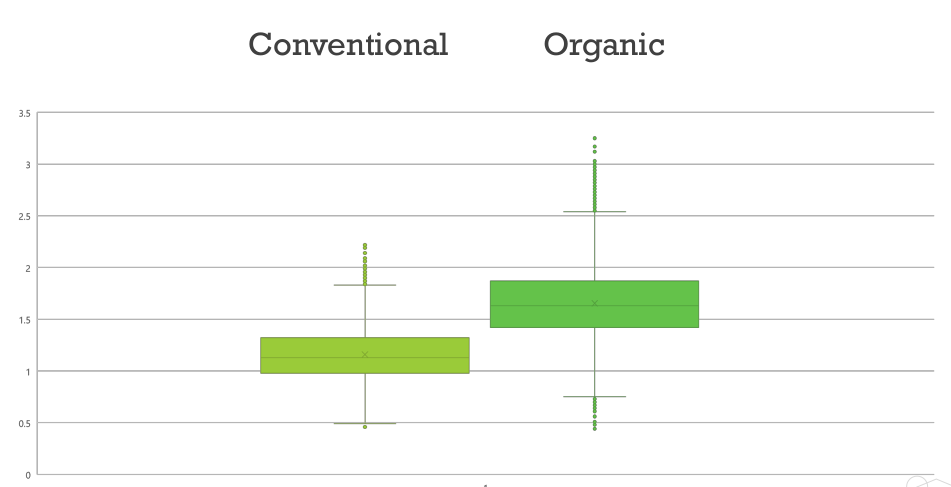
#### Fig 2.1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min | Q1 | Median | Q3 | Max | Std dev/ σ | Variance/ σ2 | CV |
| Conventional | 0.46 | 0.98 | 1.13 | 1.32 | 2.22 | 0.26304 | 0.069190 | 22.7% |
| Organic | 0.44 | 1.42 | 1.63 | 1.87 | 3.25 | 0.36350 | 0.132133 | 21.9% |

#### Fig 2.2



#### Fig 2.3



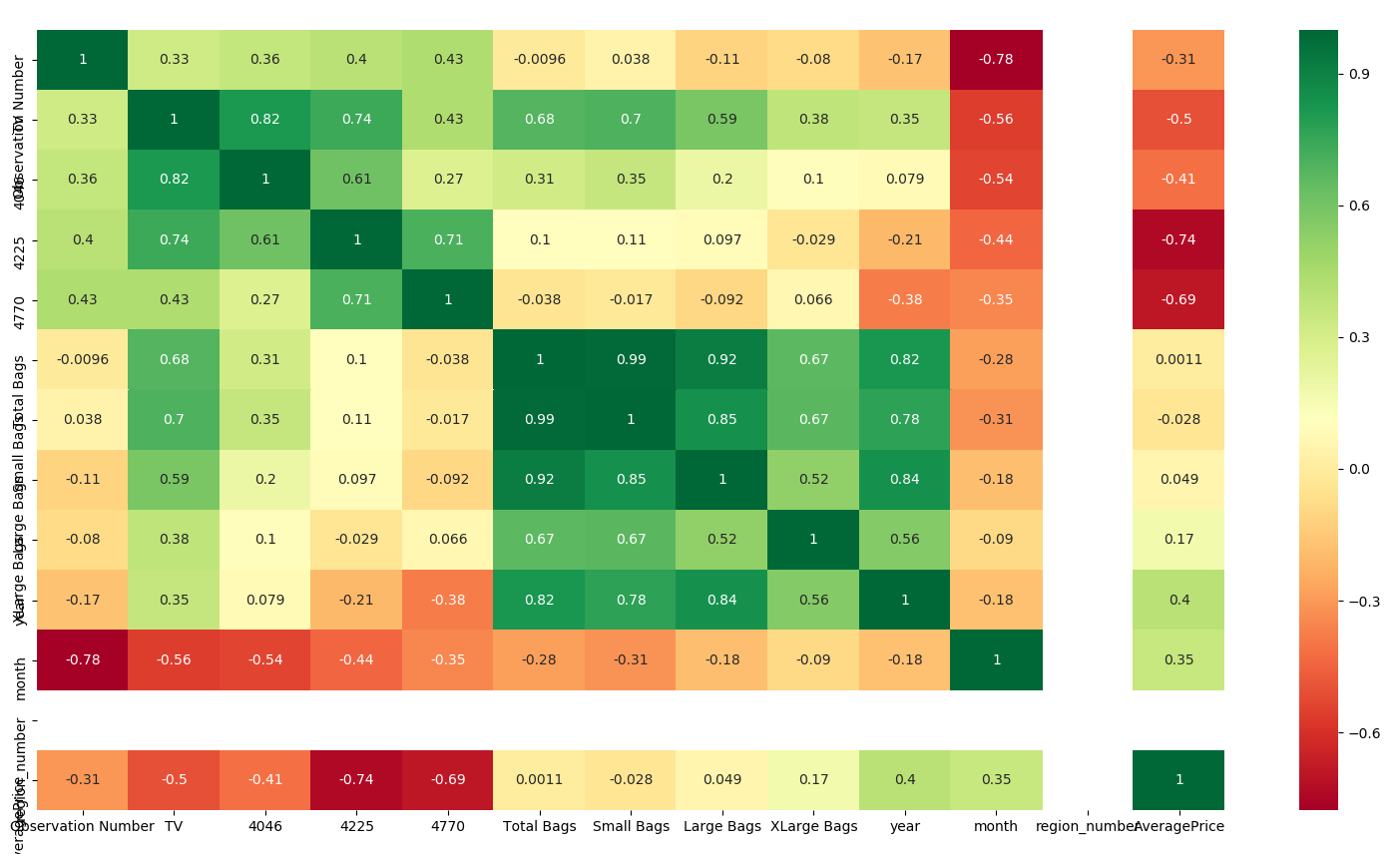
* There is a relatively large difference between the two datasets (fig 2.1). The reasons for this can be observed in the five-number summary table (fig 2.2) and the boxplot diagram (fig 2.3). The min and max values are smaller and larger respectively for the organic dataset, as opposed to the conventional dataset. The most pronounced difference between the two datasets can be observed at the max value for organic compared to conventional. This evidence has provided us with an insight as to what shape our data is.
* The IQ range of for both types indicates that the data is clustered about the mean similarly. This observation is proved to be true when we determined the standard deviation and variance, as both bell curves appear to possess similar shapes, and the coefficient of variation (CV) for both types of avocados are approximately the same.
* We can conclude that the difference between the datasets for organic and conventional avocados, is not due to how the data is clustered, but is in fact due to the presence of outliers at the maximum and minimum ends of each dataset.

### **2.3:** Feature Engineering

To make an accurate regression prediction, we decided that we would need to engineer some features. We identified the *month* as a feature we would need to add into our model as we could see that the prices tended to follow a pattern throughout the year.

We then added a numerical number feature to our categorical data in order to use it in our algorithm. In this case the categorical data that we identified that we may need to use throughout our analysis was *region\_number*. A numerical value is assigned to the corresponding region for each line in our dataset, as many machine learning algorithms cannot work with categorical data (label encoding/introducing dummy variable).

From our analysis on the dispersion of the data set, we could clearly see that there was less of a dispersion in the *conventional* type of avocados, so we decided to make our prediction using this, as the data appeared to follow a more uniform trend. We first identified our target variable (dependent variable) as *AveragePrice*, and then used this to find out which of our features had the strongest relationship with it. To do this we used a correlation matrix displayed as heatmap, as it would give us the best idea about how all our features interact with each other.



We see that the month and year that the conventional avocado was produced is highly correlated, being *0.4* and *0.35* respectively. As expected, the Total volume or *TV* is negatively correlated with *AveragePrice*(Economies of scale). Interestingly, *4225* and *4770* we’re very negatively correlated with *AveragePrice*, which was a surprise to us. Knowing this information, we then proceeded to experiment with features during the prediction stage of our project.

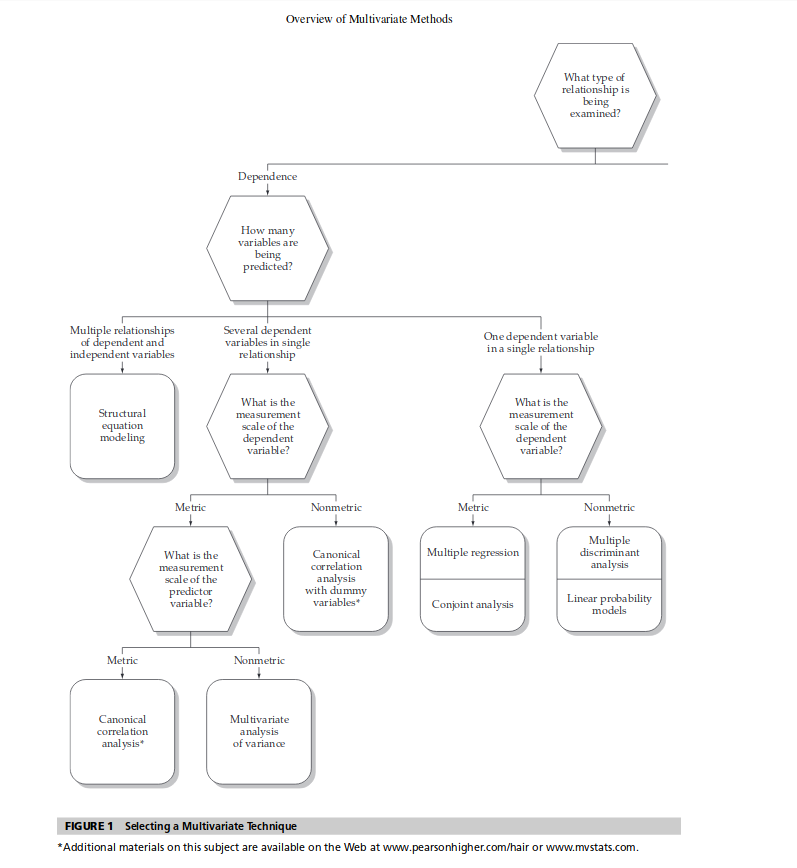
### **2.4:** Multicollinearity

We know that multicollinearity “refers to the extent to which independent variables are correlated” (StatTrek, 2019) and upon analysing the heatmap above we were concerned initially that our independent variables would suffer from multicollinearity (*Total Bags, Large Bags, Xlarge Bags)*.There was also a high degree of multicollinearity between the PLU numbers. We were initially worried that this would affect the accuracy of our prediction but after investigating this further, we assessed that the “least-squares regression equation can be highly predictive.” (Stattrek, 2019), and this is regardless of multicollinearity in our dataset. Also, none of the independent variables hold a perfect multicollinearity (with a correlation of 1 or -1) despite sources stating that if the correlations are “.75 or higher, then there may be a problem with multicollinearity”(<http://web.csulb.edu/~msaintg/ppa696/696regmx.htm>), we proceeded with the analysis.

### **2.5:** Data Standardisation

For us to achieve the best possible results from our prediction algorithm, we decided that we would standardize our data (centre the variables). It is useful for us to do this because our dataset has varying scales, for example *AveragePrice* and *TV* (Total volume). We wanted to ensure that our data was as consistent as possible, having the same format. With standardization we are rescaling our data to have a mean of 0 and a standard deviation of 1. Multiple sources recommend it in terms of numerical stability, so we felt it was best to apply it to our data set.

## Section 3: Algorithm description

The aim of our project is to predict the price (numeric), where the value is continuous and to explore the effect on different attributes on the *AveragePrice* and their ability to be able to accurately predict it. With these factors considered, we referred to the diagram below (F.Hairet.al, 2014) and were sure that multiple regression was perfect for our purposes.

### **3.1:** Multiple Linear Regression

The equation for multiple linear regression with *K* predictor variables *X1, X2,...XK* and a response *Y* is given as: (Bremer(2016))



With *B0,B1,B2,...BK* above being the Regression coefficients.

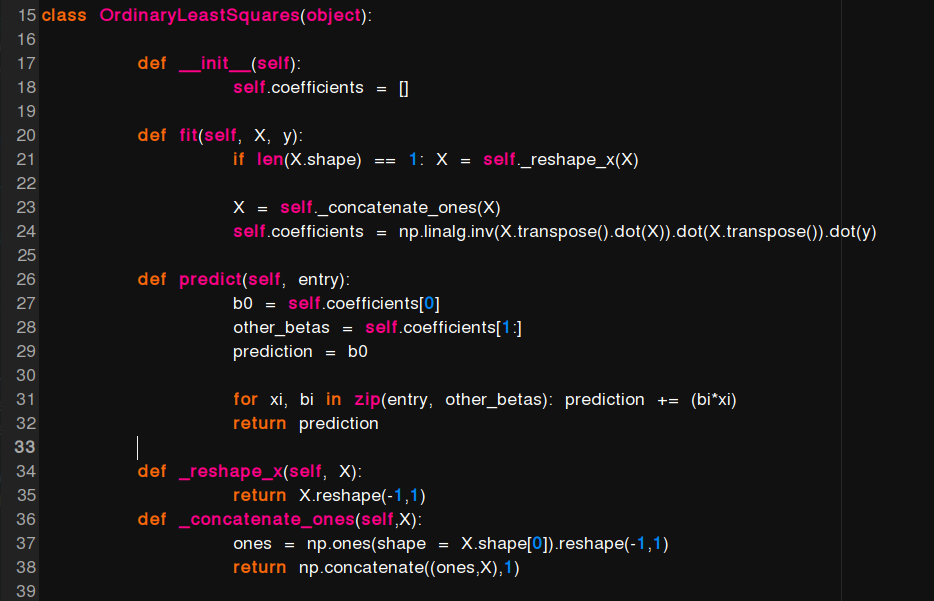
After we decided to choose Multiple Linear Regression for our price prediction, we then decided that the ordinary least squares method was the most appropriate method of multiple linear regression for our purposes. This was because of the presence of outliers, for instance a price spike in 2017 and others which were found during our data dispersion phase of our pre-processing.

OLS is known as a more rigid method of Multiple Linear Regression and therefore we hypothesised that the model would perform well regardless of these. This would also prevent the model from becoming overfitted.

### 3.2: Ordinary Least Squares algorithm

The main goal of the *ordinary least squares* method is to estimate the relationship between our independent variables and our dependent variable by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable.

Errors, when referring to Ordinary Least Squares is the difference between prediction and reality and in the case of the prediction, *AveragePrice*. In our case, our OLS equation is interested in the squares of the errors, and it will look for the perfect plane going through our data that will minimize the total sum of these squared errors, and our model will be based on the equation of that particular plane. The algorithm for the calculation in python is as followed:



## Section 4: Implementation, Results, Analysis & What we Learned

### 4.1: Splitting our dataset into Training and Test data sets

We decided that for the purposes of testing that we would split our data using a test/training split of 80/20 was ideal for our purposes.

### 4.2: Feature selection & Linear Regression Calculation

As we were aware of the curse of dimensionality (Yiu,2019), we knew that overfitting our data was a risk. Identifying a small number of meaningful features was important to us, and this would also provide the added benefit of making our prediction more efficient.

We decided to use backwards elimination to determine the optimal features to use for our predictor model. Method:

1. Select a significance level (SL). Any score greater than the SL will be removed from the model.
2. Fit model with all possible predictors.
3. Consider the predictor with the highest p-value. If the p-value is greater than the SL, proceed to step four. Otherwise the model is ready
4. Remove the predictor
5. Fit the model without this variable

We carried out this backwards elimination using the statsmodel python library as we can use the Ordinary Least Squares (OLS) class. The Significance Level is an indication that there is an X percentage likelihood that the results obtained were obtained by chance. We selected a significance level of 0.05(low enough to be satisfactory).

Initially we attempted to predict the price for avocados for all of the USA (“region” == “totalUS”), but as can be observed in fig 4.10, our predictor was too inaccurate.

We estimated the error by “computing an error based on the difference between the predicted value and the actual known value of y, for each of the test tuples X” (Roantree, 2019) Along with this estimation, we also took into account the R-squared score. The R squared score is a very valuable measure as it is a statistical measure of how close the data are to the fitted regression line.

We believe a combination of the R-squared score along with the estimation of accuracy provides us with a good understanding of how effective our predictive model is.

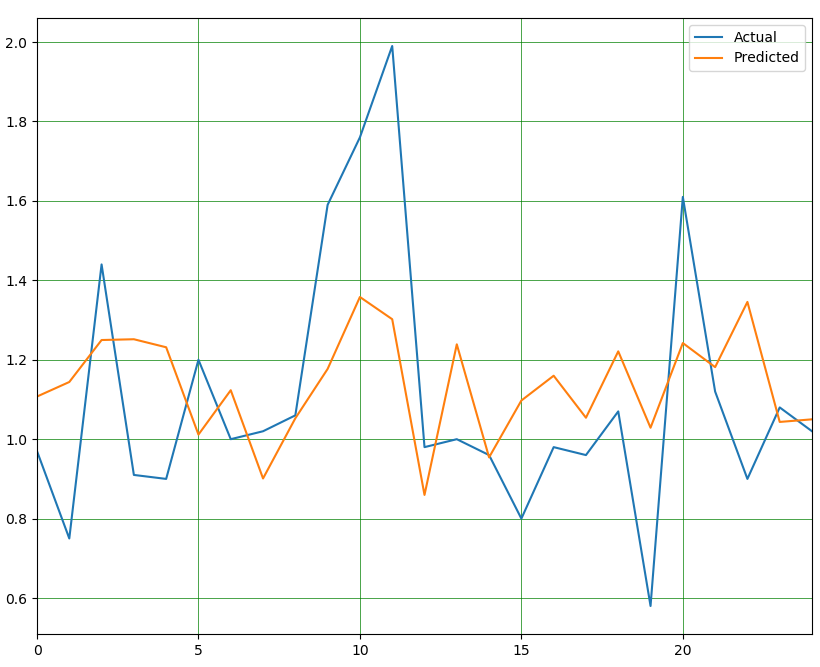
The reason we are concerned with error, is because a high R squared value is not necessarily a good reflection on the predictor. Low R squared values are not necessarily a bad reflection on the model if statistically significant features are present.

*“Total US”*

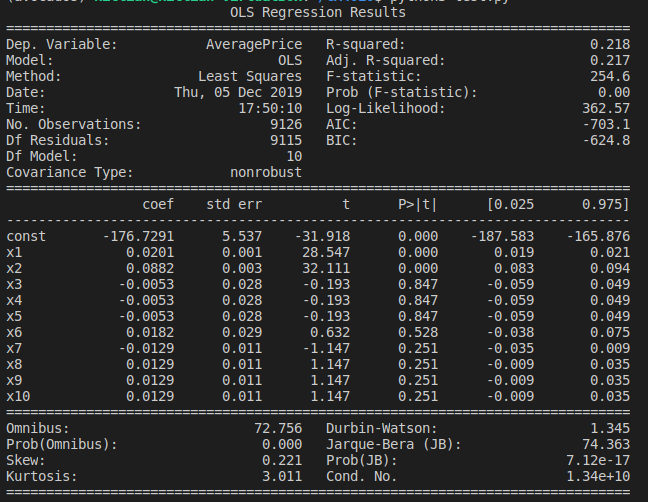
As can be observed in fig 4.11, the R squared score is 0.218. Along with this low R squared score, the error associated with this prediction is 18.1. As we have calculated a low value for R squared and a high value for error, we can refer to fig 4.10 which shows the regression line for average price. Due to these three stated calculations and observed plot, we believe that a prediction for “TotalUS” is too inaccurate. It is not surprising that our model could not accurately predict the average price across the entire US, as this is a cumulating of all regions.

We then decided that predicting the average price for individual regions could be more accurate.

#### Fig 4.10 - Plot of actual average price against predicted average price for “TotalUS” region



#### Fig 4.11 OLS Regression results for “TotalUS”

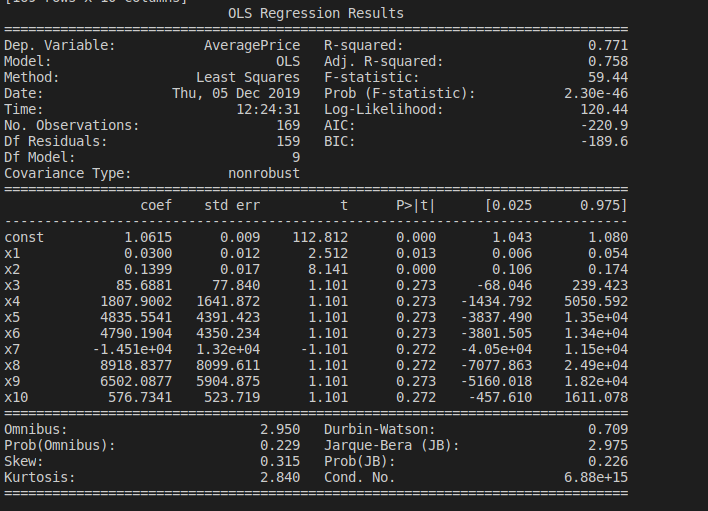


*San Diego*

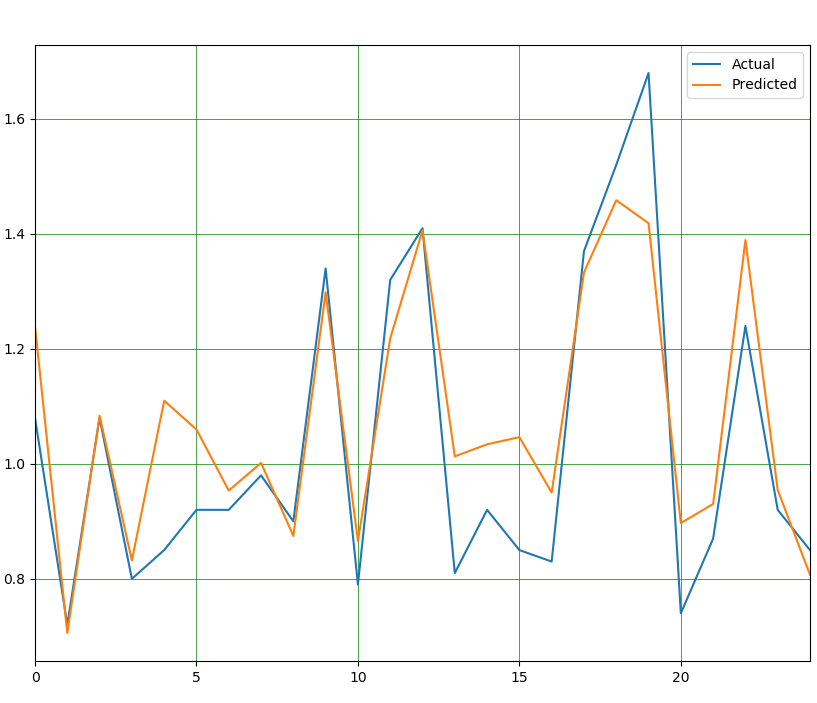
We ran our first iteration of OLS regression on the region of San Diego and with all possible predictors: month, year, XLarge Bags, Large Bags, Small Bags, Total Bags, TV (Total Volume), 4046, 4225, 4770 (For figure 4.12, the list of variables x1, x2 etc correspond to the list of predictors in a respective order). The first set of results from our OLS regression can be seen in fig 4.12.

In order to help us interpret these results, we also plotted on a graph the actual values for “AveragePrice” and the predicted values for “AveragePrice”. The error associated with our predictor with all possible features is: 9.029%. The R squared score (top right of fig 4.12) is observed at 0.771, which is factors higher than the score for total US.

#### Fig 4.12 - OLS Regression results for San Diego with all features

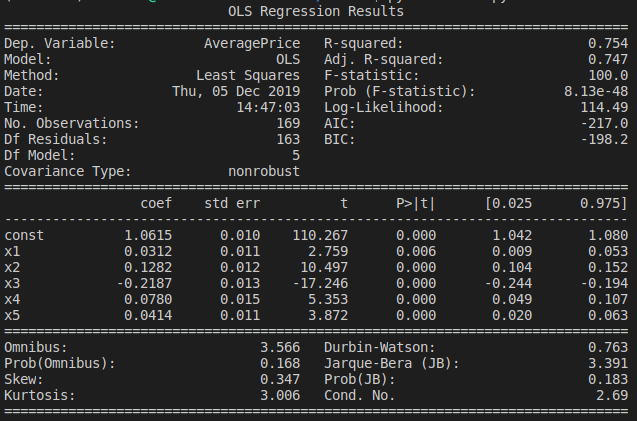


#### Fig 4.13 - Plot for San Diego of predicted average price against actual average price, with all features present

****

Following the method outlined for backward elimination, we can identify x3, x4, x5 and x9 as predictors with the highest p value. We chose to remove x9 - “4225”, as the predictor to eliminate. This process was repeated until we no more p values dropped below our SL, which indicates that our model is ready. The final list of predictors for our model are as follows: month, year, TV (Total Volume), 4046, 4770. The OLS regression results can be observed in fig 4.13.

#### Fig 4.14 - Final OLS Regression results for San Diego



We decided to carry out backwards elimination on 10 other regions because we wanted to compare the variation in features and accuracy across the Untied States and potentially identify key predictors (or lack of). We measured the error of the model before and after backwards elimination to determine the effect the removal of features had on accuracy. We also measured R squared and R squared adjusted to also measure the effect less features had. The reason we decided to investigate R squared adjusted was because R squared adjusted takes into account the number of predictors in the model, so therefore it seems an apt measure of how successful our backwards elimination is.

#### Figure 4.15 - Table containing 10 random US regions, along with R-squared scores and features remaining after backward elimination (BE)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Region** | **R2 before BE** | **R2 after BE** | **Adjusted R2 before BE** | **Adjusted R2 after BE** | **Error before** | **Error after** | **Features remaining** |
| San Diego | 0.771 | 0.754 | 0.758 | 0.747 | 9.029301859668255% | 9.387985573629196% | Month, year, TV, 4046, 4770 |
| New York | 0.636 | 0.619 | 0.613 | 0.612 | 8.488170064992378% | 8.653530104527624% | Year, XLarge Bags, 4225 |
| Jacksonville | 0.731 | 0.721 | 0.714 | 0.712 | 8.765385299293055% | 8.372218821411852% | Month, year, TV, 4046, 4225 |
| Baltimore Washington | 0.696 | 0.691 | 0.679 | 0.682 | 8.652739759729277% | 8.37129663225907% | year , large bags, small bags, Total bags, 4225 |
| Houston | 0.648 | 0.637 | 0.628 | 0.628 | 10.23790240356025% | 9.394185290612725% | Year, Total bags, TV, 4225 |
| Greatlakes | 0.641 | 0.634 | 0.619 | 0.621 | 9.688167758564745% | 9.3300775872665% | Year, Large bags, Small bags, Total bags, TV, 4046 |
| Orlando | 0.754 | 0.75 | 0.738 | 0.741 | 6.905606733441062% | 6.932061632635289% | Month, year, total bags, TV, 4046, 4225 |
| South Carolina | 0.593 | 0.579 | 0.567 | 0.564 | 9.230447529162241% | 9.040039004753629% | Year, Large bags, small bags, Total bags, 4046, 4225 |
| Nashville | 0.607 | 0.6 | 0.583 | 0.582 | 11.525443619098915% | 9.46872297164797% | Month, year, small bags, total bags, TV, 4046, 4225 |
| Pittsburgh | 0.599 | 0.598 | 0.579 | 0.583 | 5.957601761210877% | 5.93780282627665% | Month, XLarge bags, Large bags, TV, 4046, 4225 |

### **4.3** What we learned whilst carrying out these experiments

#### 4.3.1 Analysis of backwards elimination across 10 different regions:

##### 4.3.1.1 Features:

From figure 4.15, there are no common features that occur in all 10 of the randomly selected regions. “Year” occurs the most, 9 out of 10 times. The next most common feature is “4225” occurring 8 out of 10 times, and the next is “TV” (Total Volume) and “4046”, 7 out of 10 times. It is also observable that there is no consistent combination of features that occur alongside each other in each region. So although there are four features that occur 9, 8 and 7 times, there is no pattern at which they occur.

##### 4.3.1.2 Error:

Across all ten regions, error ranges from 5 - 11 %. It is clear from fig 4.15 that the removal of predictors has an effect on the error, either positively or negatively. Generally, the variation in error is relatively small (typically ~ < 1 %). The removal of predictors generally decreases the error, as 7 of the 10 regions’ error decreases when selected features are removed. The largest decrease in error can be seen in Nashville, at 2 percent. Therefore, from our 10 random regions we can observe that the removal of predictors has not had a net negative effect on the accuracy of our model.

##### 4.3.1.3 R squared and R squared adjusted:

A decrease in R squared when predictors are removed is the general trend observed from our 10 sample regions, but the decrease is relatively low (~ < 0.02). Because R squared cannot decrease when features are added, the best expectation for this value is to remain the same when features are removed. But as stated, the decrease is very small.

As mentioned previously, R squared adjusted is a more accurate measure for backwards elimination as this value does account for the number of features. Across all ten regions, 4 R squared adjusted values increased when features were removed, 1 region’s value remained the same. Of the four regions that R squared adjusted value decreased, the largest observable decrease was 0.01, and the smallest was 0.001.

##### 4.3.1.4 Combining all the above:

Regarding our predictor models, backwards elimination was effective as we were able to reduce the number of features in each region, while maintaining (and sometimes improving) accuracy, R squared and R squared adjusted, while also removing multicollinearity from the features.

Considering the presence of features, accuracy and R squared and adjusted, the accuracy of our predictor model for conventional avocados varies across different regions. The lack of consistent features or a consistent combination of features indicates that with the available features in this dataset, that the average price for avocados is independent from region to region. With this dataset, it is possible to provide a relatively accurate predictor model (as we have demonstrated), for each individual region. But as previously mentioned, the statistical significance of features varies from region to region. We believe this is reflected by the lack of statistically significant features for the “totalUS”.

### 4.3.2 Conclusions

From carrying out this project we learnt about the difficulties of linear regression. We learned techniques and the importance of feature engineering. We learned about multicollinearity and how it affects regression, and data standardization and the pitfalls and benefits of it and we learned how to check if a dataset is clean and how to prepare our dataset for regression.

In terms, of our prediction we learned in depth about multiple linear regression and OLS, and the pitfalls and benefits of this. We learned about selecting training and test splits and backwards elimination as a method of selecting features, and why we should be sceptical about this.

With the results of our predictions, probably the most important thing that we have learned is how to interpret our results. We developed, as the project went on, a certain level of cynicism towards positive results and predictions that we believe is incredibly beneficial. We also began to interpret the meaning and goal of our project differently as it progressed, as the goal originally was predict price X at date X, but as we learned more about MLR, we became more interested in what factors can help us to accurately develop a model for predicting the price of an avocado, and what factors cannot help us.

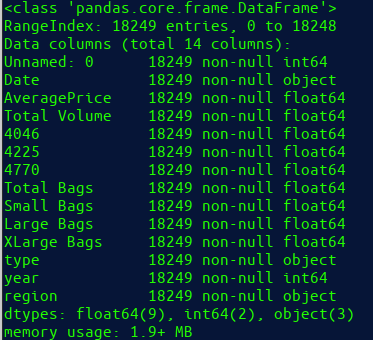
All in all, this project presented us with many challenges but also taught us a lot about data science and regression. We will now plan to use that knowledge going forward with the year and in our professional lives.

## Section 5: References

* <https://hassavocadoboard.com/inside-hab/>
* <https://www.researchgate.net/publication/237641203_A_History_of_the_Avocado_Industry_in_California>
* <https://www.kaggle.com/neuromusic/avocado-prices>
* Ferman(2015) <https://www.washingtonpost.com/news/wonk/wp/2015/01/22/the-sudden-rise-of-the-avocado-americas-new-favorite-fruit/>
* CityMayors, 2019 <http://www.citymayors.com/gratis/uscities_100.html>
* <https://towardsdatascience.com/regression-predict-house-price-lesson-2-5e23bec1c09d>
* Yiu, (2019) <https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e>
* <https://stattrek.com/multiple-regression/multicollinearity.aspx> (2019)
* <https://loop.dcu.ie/pluginfile.php/2829664/mod_resource/content/2/1-Classification-handouts.pdf>
* 9 781292 021904ISBN 978-1-29202-190-4Multivariate Data AnalysisJoseph F. Hair Jr. William C. BlackBarry J. Babin Rolph E. AndersonSeventh Edition 2014
* <https://medium.com/@erika.dauria/looking-at-r-squared-721252709098>
* <http://web.csulb.edu/~msaintg/ppa696/696regmx.htm>
* CA4010 course notes: classification slides 1.12 - [**https://loop.dcu.ie/pluginfile.php/2829664/mod\_resource/content/2/1-Classification-handouts.pdf**](https://loop.dcu.ie/pluginfile.php/2829664/mod_resource/content/2/1-Classification-handouts.pdf)).
* [**https://loop.dcu.ie/pluginfile.php/2829664/mod\_resource/content/2/1-Classification-handouts.pdf**](https://loop.dcu.ie/pluginfile.php/2829664/mod_resource/content/2/1-Classification-handouts.pdf) Slide 12
* Bremer(2016)<http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf>.

**Code snippets:**

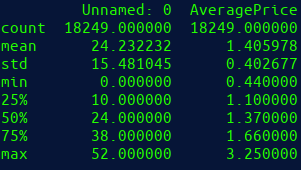
**1)**

****

**2)**



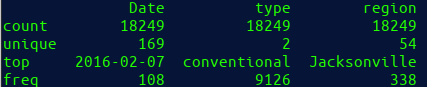
3)

****

**4)**



5)



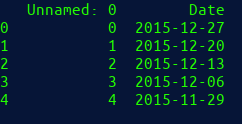
6)



7)



8)



9)



10)

