Avocado Market Behaviour analysis using the Historical data on avocado prices and sales volume in multiple US markets

Data Preparation and Statistical Techniques
Higher Diploma in Science in Data Analytics for Business
Aldana Louzan and Noel Cosgrave

By Conor Sheehan (sba 20128), Eric Parfrey (sba 20129),
Giuliano Silva (sba 20130) and Hasan Aziz (sba 20140)

CCT College Dublin

Table of Contents

1. Business Understanding

The avocado as a foodstuff has grown in popularity over the previous 20 years in Western diets due to an increased global interest in Latin American foods with which it is associated but also as a result of it being associated with healthier diets. This product was quickly recognized among a fruit that has several benefits and it was no coincidence that its market in the US began to grow and become profitable and expansive. This market has extreme potential because it involves a consumer who needs to eat the product on a daily basis in order to obtain the benefits that this fruit offers. After initial analysis, we decided that the dataset found in the American consumer market of avocados would be worthy of further investigation with particular focus on region and type.

1.2 Data Selection

The CSV file that our group has chosen contains data from the US market from 2015 to 2018 compiled as a result of Hass Avocado sales scans and represents retail volume (units) by US regions, type, size and price. The dataset was downloaded from the Kaggle website and contains 18,249 rows and 14 columns and non-missing values. One of the requirements of this project is to have a minimum of 15 variables and on initial inspection they are not present but we will work on the Data Preparation and Feature Engineering by creating new features in order to facilitate a number equal to or greater than 15.

Data Dictionary			
Data Item	**Non-null count**	**Data Type**	**Description**
Unnamed:0	18,249	int	Index
Date	18,249	object	The date of the observation
AveragePrice	18,249	float	Represents a per unit cost in US dollar

Total Volume	18,249	float	Total number of avocados sold
4046	18,249	float	Total number of avocados with PLU 4046 sold (Small/Medium Hass Avocado (3-5oz avocado)
4225	18,249	float	Total number of avocados with PLU 4225 sold (Large Hass Avocado (8-10oz avocado))
4770	18,249	float	Total number of avocados with PLU 4770 sold (Extra Large Hass Avocado (10-15oz avocado))
Total Bags	18,249	float	Total number of bags
Small Bags	18,249	float	Size of the bag
Large Bags	18,249	float	Size of the bag
XL Bags	18,249	float	Size of the bag
Type	18,249	object	Conventional or organic
Year	18,249	int	Self-explained
Region	18,249	object	The city or region of the observation

Table 1. Data Dictionary.

2. Data Understanding

Our first impression was that the dataset chosen meets all the criteria of this assignment. Another discovery that was made at the early stage was that the avocado has three types of size. The term PLU refers to Product Lookup Codes and only refers to Hass avocados:

- Small/Medium Hass Avocado (3-5oz avocado)
- Large Hass Avocado (8-10oz avocado)
- Extra Large Hass Avocado (10-15oz avocado)

The average price will be the target variable on which we will apply Regression techniques. Second, the variable type which can be used as our categorical variable will be transformed by means of Dummy Encoding which categorizes values into 0 and 1. By doing that we will be able to use the Binary Logistic Regression model. We counted the values for these two types as shown in Table 2 and we learned that the amount of types of products present in this dataset seem to be equally distributed.

Avocado types	
Type	**Non-null count**
Conventional	9,126
Organic	9,123

Table 2. Avocado types.

The first 5 rows of the Avocado dataset can be seen in Figure 1 below. Before we start off the analysis we need to fix few things:

- Create a column 'Month' from the 'Date'
- Rename the columns to an appropriate name following the Table 1

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

Figure 1. Avocado dataset.

Figure 2 shows the results:

	Unnamed: 0	Date	AveragePrice	Total Volume	Small_Medium	Large	ExtraLarge	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region	month
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany	12
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany	12
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany	12
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany	12
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany	11

Figure 2. Avocado dataset 1.

We discovered that there were a total number of regions of 54. We removed the region 'TotalUS' so that it would not feature in further analysis. In doing so , we removed 338 rows.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17911 entries, 0 to 18248
Data columns (total 15 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    17911 non-null  int64
 1   Date          17911 non-null  object
 2   AveragePrice  17911 non-null  float64
 3   Total Volume  17911 non-null  float64
 4   Small_Medium  17911 non-null  float64
 5   Large         17911 non-null  float64
 6   ExtraLarge    17911 non-null  float64
 7   Total Bags    17911 non-null  float64
 8   Small Bags    17911 non-null  float64
 9   Large Bags    17911 non-null  float64
 10  XLarge Bags   17911 non-null  float64
 11  type          17911 non-null  object
 12  year          17911 non-null  int64
 13  region        17911 non-null  object
 14  month         17911 non-null  int64
dtypes: float64(9), int64(3), object(3)
memory usage: 2.2+ MB
```

Figure 3. Data types after initial clean.

2.1 Exploration Data Analysis

Our target variable is positive skewed as Figure 4 shows below:

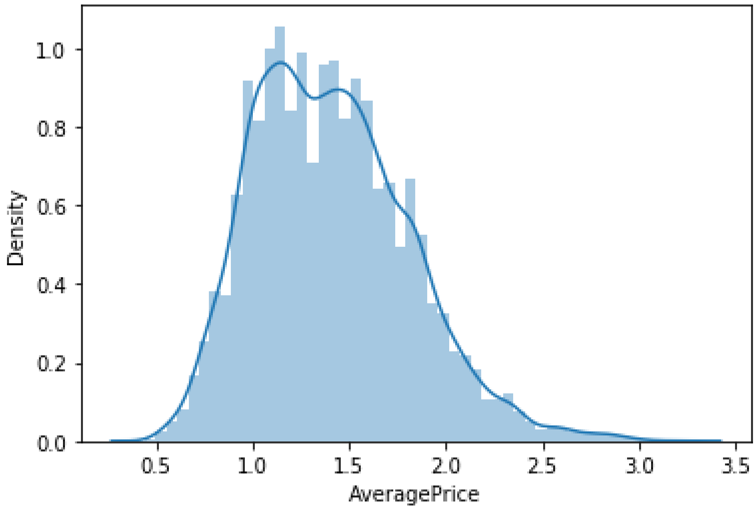

Figure 4. Average price distribution.

Figure 5 below shows the comparison of average price of the two types of avocado. It also shows that the demand for conventional avocados is higher than the organic type and also that the range of average price is higher for organic than conventional. They present an area in which they intercept which is approximately between 1.0 and 2.0 which lies within the range of price upon which they compete with each other.

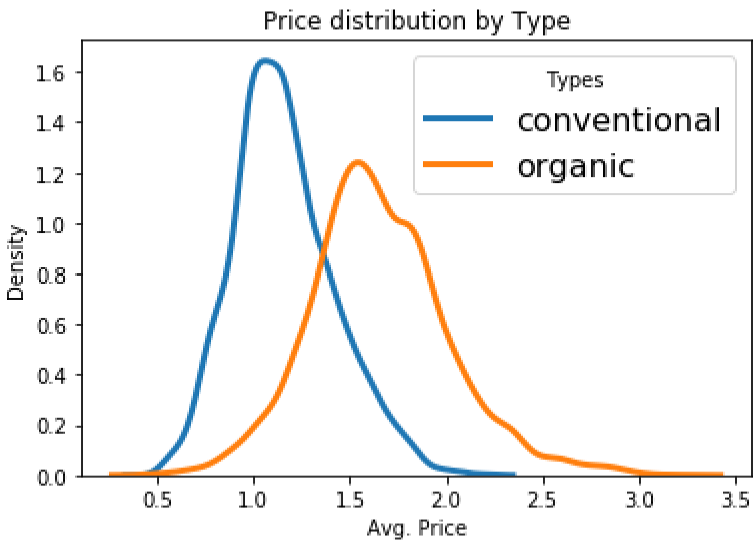

Figure 5. Price distribution by type.

Figure 6 presents to us the average price of the top 5 regions ordered from the highest price to the lowest:

region	AveragePrice
HartfordSpringfield	1.818639
SanFrancisco	1.804201
NewYork	1.727574
Philadelphia	1.632130
Sacramento	1.621568

Figure 6. Price by region.

Figure 7 shows a barplot with all the regions included:

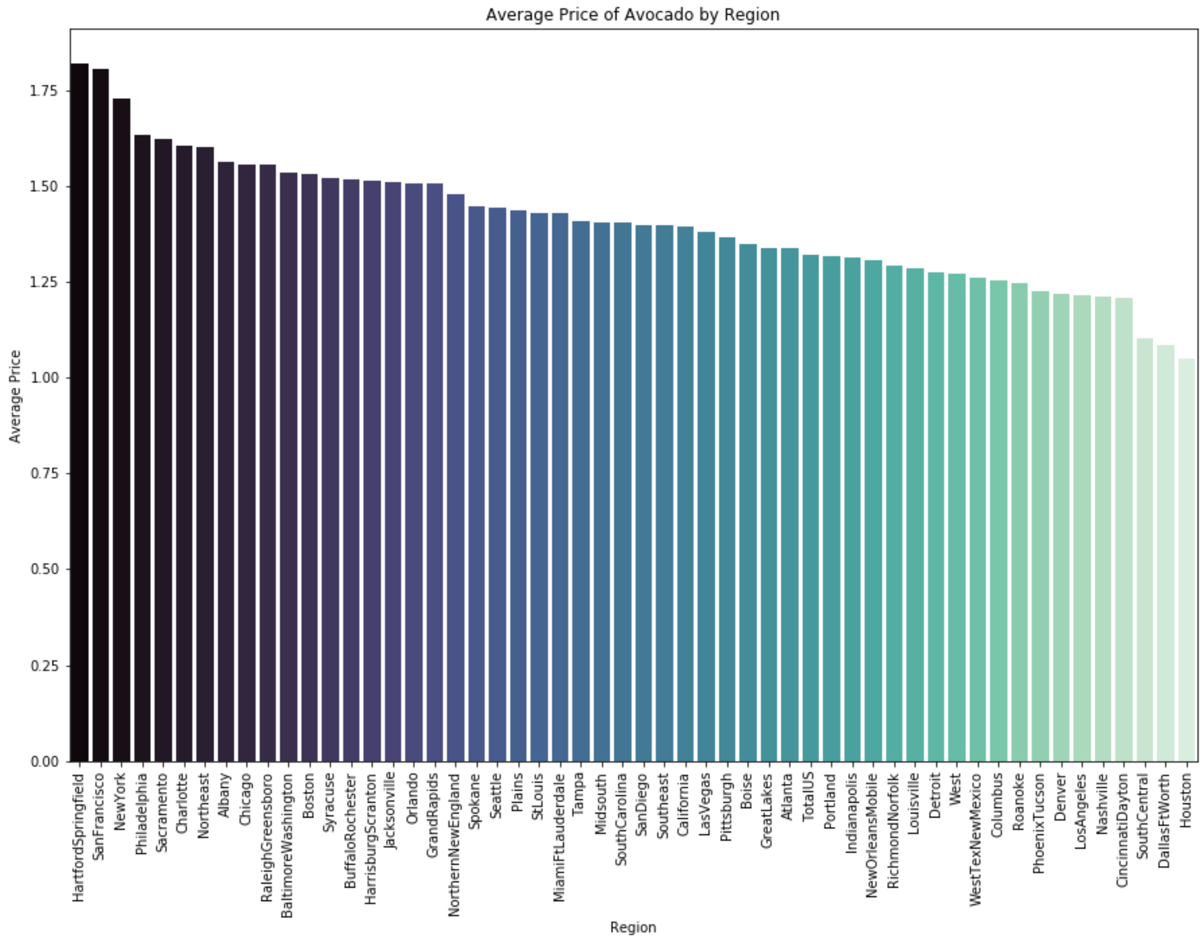

Figure 7. Barplot price by region.

The average price of an organic avocado is more expensive than the conventional as expected and is shown in Figure 5:

region	type	AveragePrice
HartfordSpringfield	organic	2.229231
SanFrancisco	organic	2.211243
NewYork	organic	2.053018
Sacramento	organic	1.969172
Charlotte	organic	1.936982

Figure 8. Price and type by region.

Figure 9 shows that the average price of a conventional avocado varies more than the organic type. The first months showed a lot of volatility in the price of the conventional variety however from the last half of May until the beginning of September they shared similar growth.

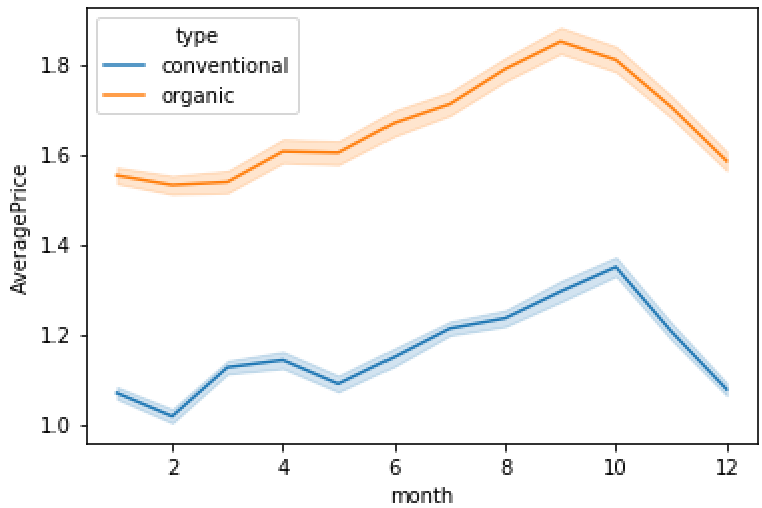

Figure 9. Linechart price and month by type.

Throughout the years the total volume from 2015 increased considerably when compared to 2018.

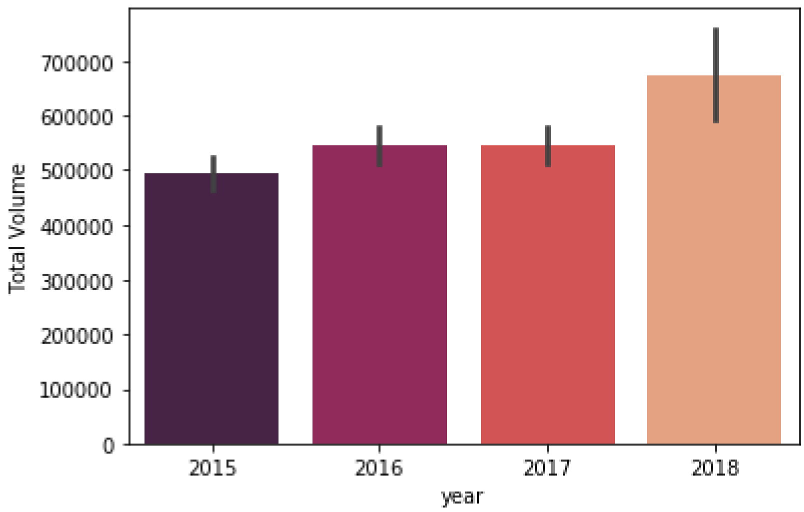

Figure 10. Barchart total volume vs year.

The total number of avocados sold within the US between 2015 and 2018 can be seen above in Figure 11. The West coast population seems to be the biggest consumer of the fruit.

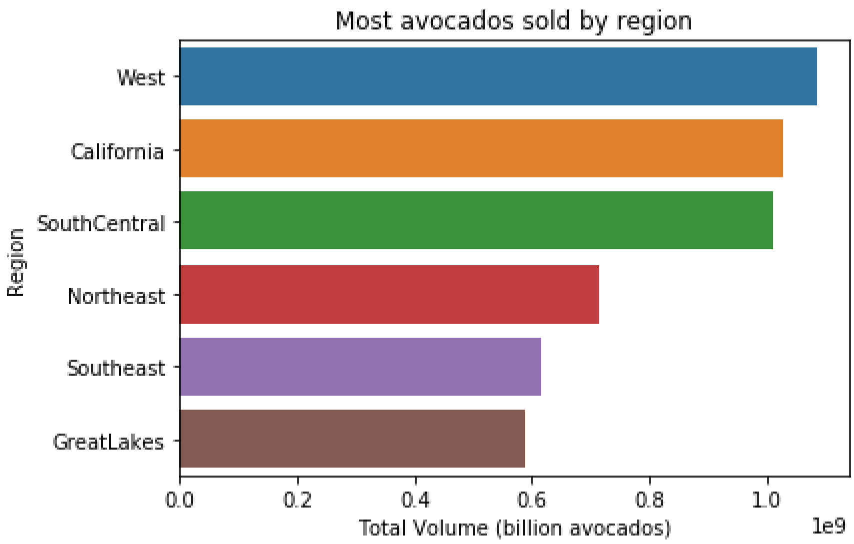

Figure 11. Horizontal bar region vs total volume.

Figure 12 shows the relationship between 'Price' and 'Total Volume' it seems ‘Price’ follows a negative correlation.

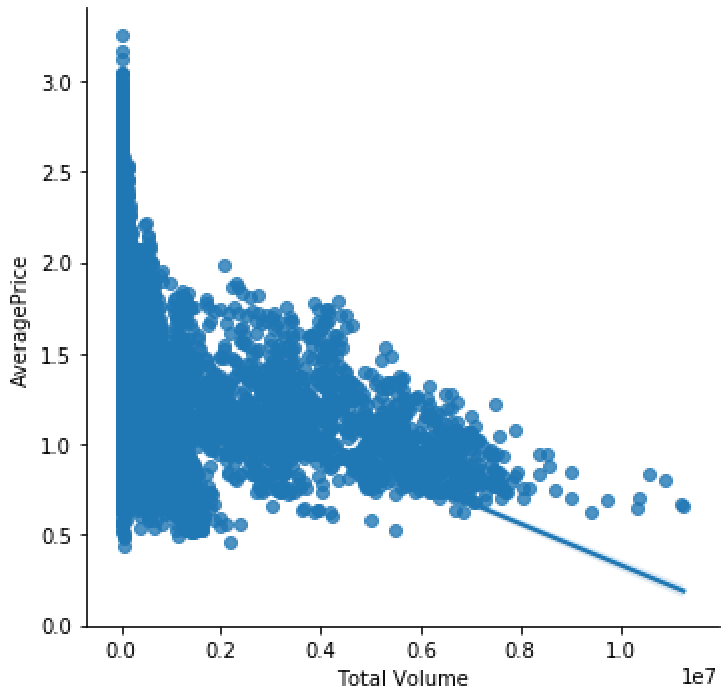

Figure 12. Scatterplot price vs total volume.

2.2 Descriptive Statistics

Table 3 shows the data types in our dataset:

Variables types	
Type	**Count**
Numerical	12

Categorical	3

Table 3. Variables types.

All the descriptive statistical metrics for numerical variables are present in Figure 4. From the mean, standard deviation and quartiles we are able to figure out how the data is distributed. We can easily see that the minimum section has a lot of zeros which will not contribute to our model as a result of it being able to affect the residuals of the regression.

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	17911.0	24.232148	1.548100e+01	0.00	10.000	24.00	38.000	52.00
AveragePrice	17911.0	1.407619	4.042530e-01	0.44	1.100	1.37	1.670	3.25
Total Volume	17911.0	539258.690838	1.224332e+06	84.56	10571.020	100154.13	400176.680	11274749.11
Small_Medium	17911.0	183807.409290	5.151059e+05	0.00	819.660	7824.43	101488.815	5160896.68
Large	17911.0	188223.112232	4.519856e+05	0.00	2909.610	26701.99	131755.215	5402444.45
ExtraLarge	17911.0	14551.234381	4.881754e+04	0.00	0.000	164.23	5736.735	804558.25
Total Bags	17911.0	152675.731028	3.645992e+05	0.00	4905.195	37551.02	103691.600	4145406.70
Small Bags	17911.0	116202.868898	2.787596e+05	0.00	2700.335	24530.62	79282.590	3403581.49
Large Bags	17911.0	34505.693530	1.139477e+05	0.00	112.995	2459.22	19421.705	2838239.39
XLarge Bags	17911.0	1967.168041	8.186402e+03	0.00	0.000	0.00	106.760	131300.76
year	17911.0	2016.147898	9.399389e-01	2015.00	2015.000	2016.00	2017.000	2018.00
month	17911.0	6.177210	3.534132e+00	1.00	3.000	6.00	9.000	12.00

Figure 13. Central Tendency Measures.

Figure 14 shows all the numerical variables plotted:

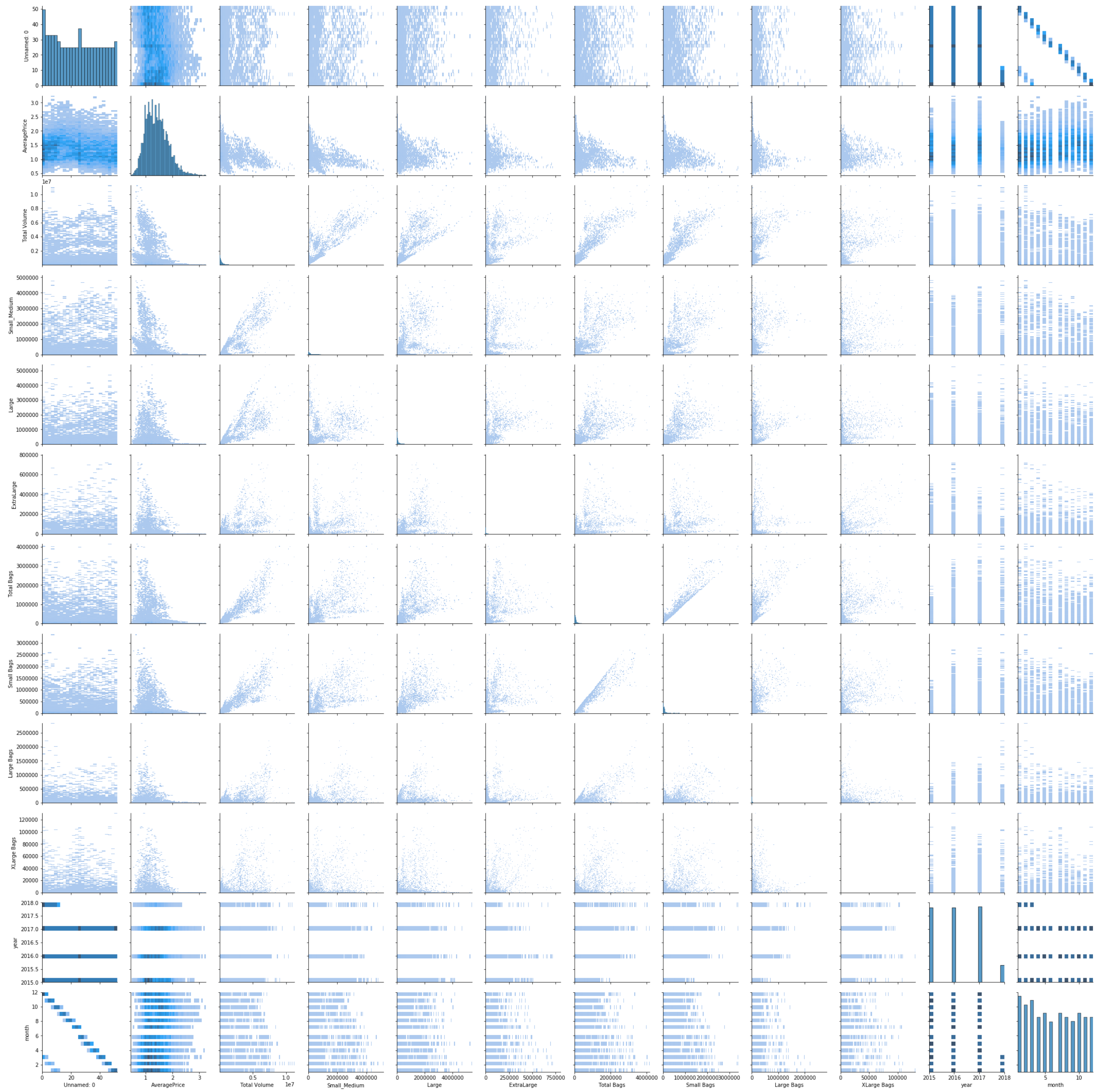

Figure 14. Pairplot numerical variables.

Figure 15 shows the distribution of avocado size independent variables and they seem to positive skewed:

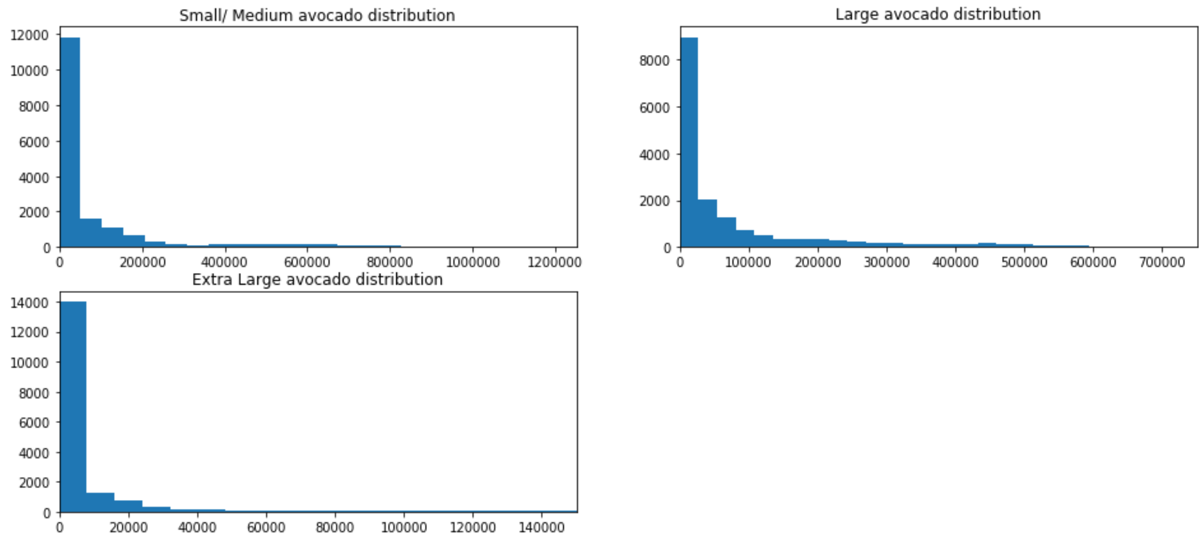

Figure 15. Avocado size distribution.

The same occurs in relation to the size of bags in Figure 16.

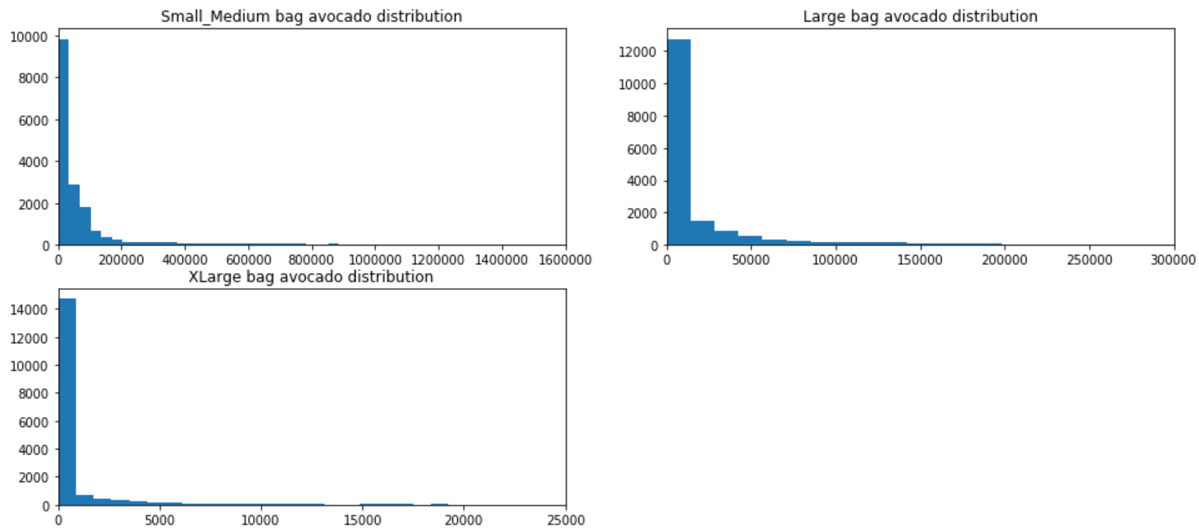

Figure 16. Avocado bags distribution.

Our target variable is average price so the first thing is to check which independent variable has a strong correlation to it. Figure 17 below shows the correlation of all numerical features. All the dependent variables seem to have a weak to moderate correlation with price.

	AveragePrice	Small_Medium	Large	ExtraLarge	Small Bags	Large Bags	XLarge Bags	year	month
AveragePrice	1.000000	-0.342105	-0.267643	-0.241213	-0.296151	-0.248909	-0.154424	0.091897	0.161463
Small_Medium	-0.342105	1.000000	0.603442	0.509280	0.761604	0.589649	0.436249	0.004244	-0.039269
Large	-0.267643	0.603442	1.000000	0.623368	0.782892	0.466107	0.449903	-0.015449	-0.037382
ExtraLarge	-0.241213	0.509280	0.623368	1.000000	0.566304	0.343136	0.587963	-0.050252	-0.046055
Small Bags	-0.296151	0.761604	0.782892	0.566304	1.000000	0.613817	0.587470	0.108639	-0.039521
Large Bags	-0.248909	0.589649	0.466107	0.343136	0.613817	1.000000	0.267308	0.118942	-0.027787
XLarge Bags	-0.154424	0.436249	0.449903	0.587963	0.587470	0.267308	1.000000	0.110051	-0.017466
year	0.091897	0.004244	-0.015449	-0.050252	0.108639	0.118942	0.110051	1.000000	-0.177048
month	0.161463	-0.039269	-0.037382	-0.046055	-0.039521	-0.027787	-0.017466	-0.177048	1.000000

Figure 17. Numerical variables correlation.

Focusing on the categorical variables Table 5 summaries all the information such as cardinality and mode.

Categorical variables metrics		
Data Item	**Cardinality**	**Mode**
Region	53	Albany
Type	2	Conventional

Table 4. Categorical variables metrics.

From Figure 18 we can visualize Table 2 via a barplot. This will be our variable for the classification task. Once the variable is encoded via a binary method the return will be a value that results in 0 or 1. In doing so, we can switch our predictions to see if our model can predict whether this fruit was conventional or organic. It's important that they are balanced otherwise it could result in a biased model.

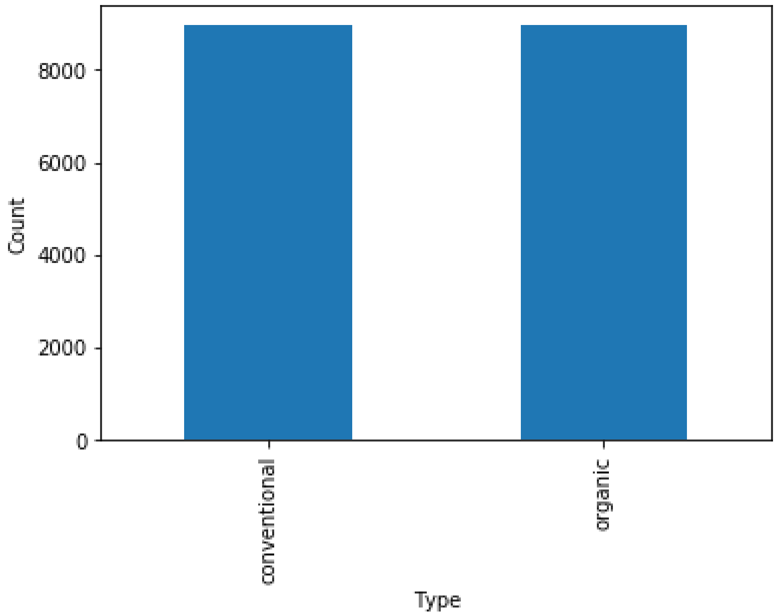

Figure 18. Barplot count of type.

Figure 19 shows us the distribution of regions.

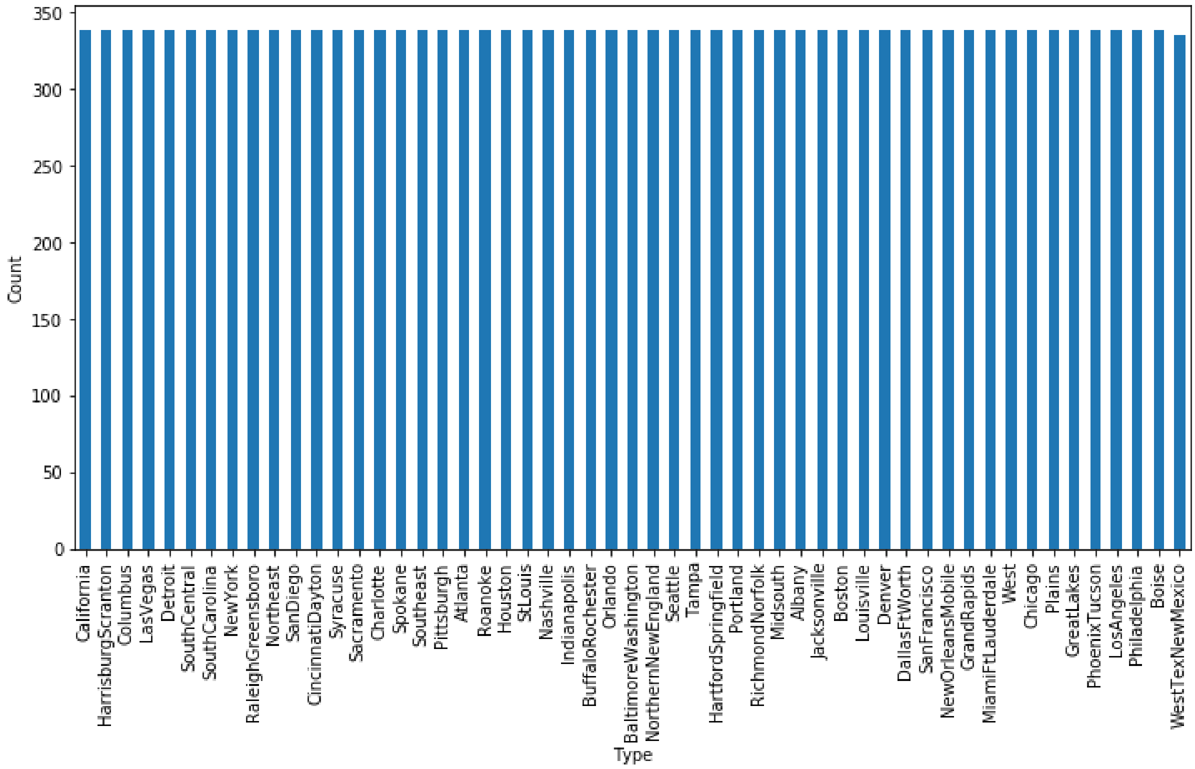

Figure 19. Barplot count of regions.

2.3 Chi-square test

The objective is to test whether the average price for each year is independent at a superiority level of 5%. The results of which are shown in Figure 20 & 21 below.

- The null hypothesis: Average prices are independent.
- Alternative hypothesis: Average price are dependent

year	2015	2016	2017	2018
type				
conventional	1.079198	1.106705	1.296269	1.129167
organic	1.676552	1.573407	1.737107	1.567421

Figure 20. Chi-square test average price by type and year.

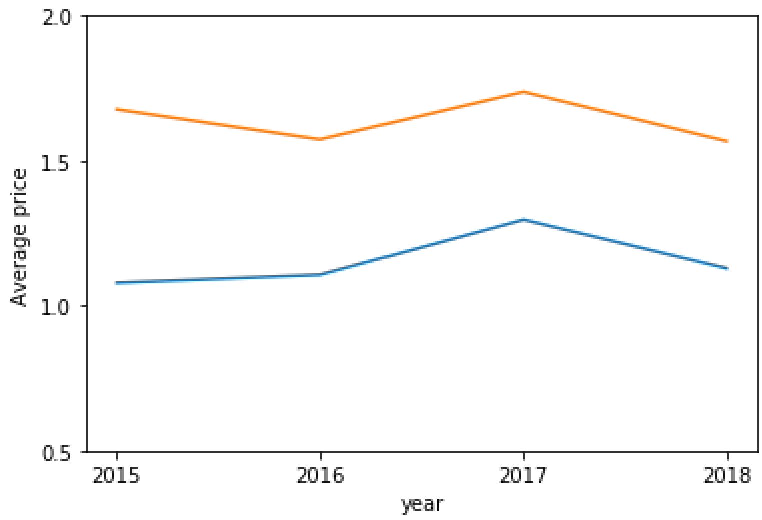

Figure 21. Chi-square line chart average price x year.

The resulting p-value is 99% which means that we do not reject the null hypothesis at the 95% level of confidence.

year	2015	2016	2017	2018
region				
Albany	1.538750	1.533942	1.637830	1.435833
Atlanta	1.380577	1.214135	1.428774	1.288750
BaltimoreWashington	1.368846	1.587596	1.679434	1.378333
Boise	1.373750	1.141923	1.492642	1.492500
Boston	1.473558	1.426154	1.679528	1.576667

Figure 22. Chi-square test average price, type, year and region.

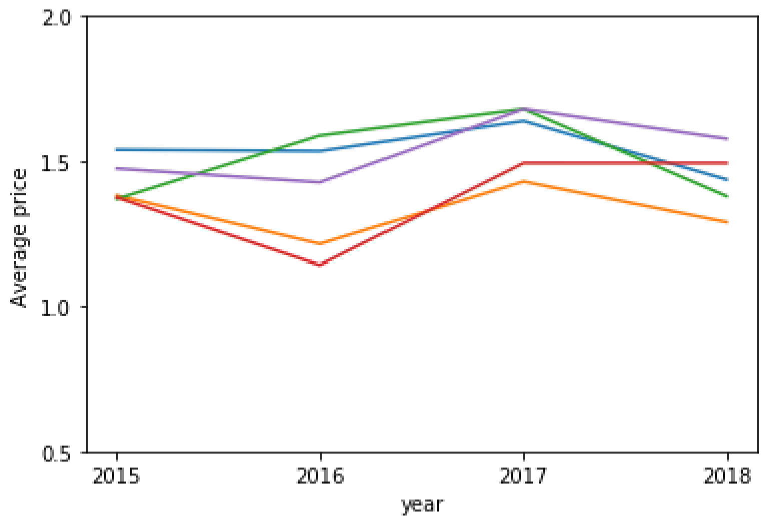

Figure 23. Chi-square line chart 2 average price x year.

The p-value is found to be 100% which means that we do not reject the null hypothesis at the 95% level of confidence.

2.4 ANOVA test one way

Our objective is to test the average price of 2015, 2016, 2017, 2018 at a superiority level of 5%

- The null hypothesis: The average variance for each year is the same.
- Alternative hypothesis: The average variance for each year is different

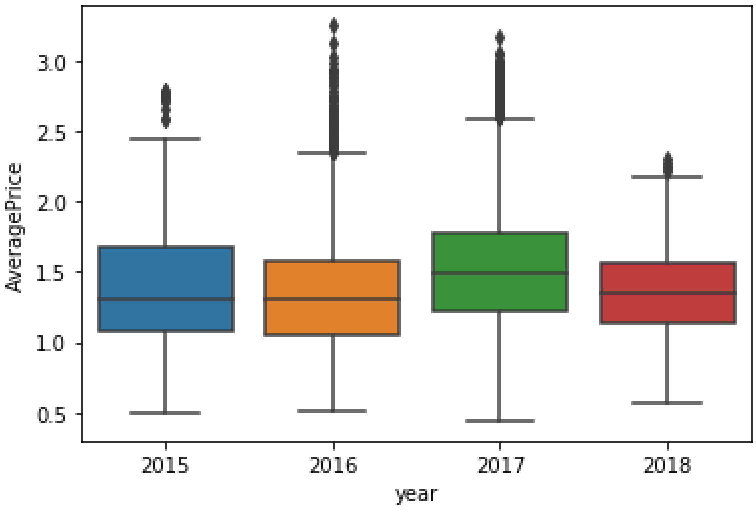

Figure 24. Boxplot price vs year.

The results of the ANOVA test one way is a p-value less than 5% and as a result, we can reject the null hypothesis.

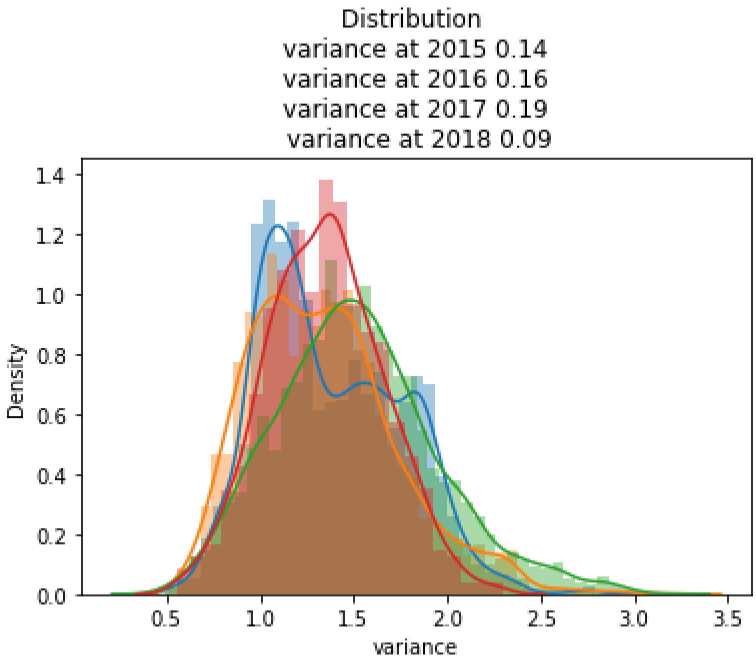

Figure 25. Variance of all years.

2.4 ANOVA test two ways

Our objective is to test the average price of 2015, 2016, 2017, 2018 versus type (conventional and organic) at a superiority level of 5%

- The null hypothesis: The average variance for each year is the same.
- Alternative hypothesis: The average variance for each year is different

Figures 26 and 27 resume the p-value for each interaction:

OLS Regression Results

Dep. Variable:	AveragePrice	**R-squared:**	0.419
Model:	OLS	**Adj. R-squared:**	0.419
Method:	Least Squares	**F-statistic:**	1846.
Date:	Sun, 03 Jan 2021	**Prob (F-statistic):**	0.00
Time:	10:58:43	**Log-Likelihood:**	-4325.9
No. Observations:	17911	**AIC:**	8668.
Df Residuals:	17903	**BIC:**	8730.
Df Model:	7		
Covariance Type:	nonrobust		

Figure 26. OLS Regression results.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0792	0.006	183.861	0.000	1.068	1.091
C(type)[T.organic]	0.5974	0.008	71.956	0.000	0.581	0.614
C(year)[T.2016]	0.0275	0.008	3.314	0.001	0.011	0.044
C(year)[T.2017]	0.2171	0.008	26.274	0.000	0.201	0.233
C(year)[T.2018]	0.0500	0.014	3.686	0.000	0.023	0.077
C(type)[T.organic]:C(year)[T.2016]	-0.1307	0.012	-11.129	0.000	-0.154	-0.108
C(type)[T.organic]:C(year)[T.2017]	-0.1565	0.012	-13.394	0.000	-0.179	-0.134
C(type)[T.organic]:C(year)[T.2018]	-0.1591	0.019	-8.299	0.000	-0.197	-0.122

Omnibus:	721.577	Durbin-Watson:	0.331
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1410.742
Skew:	0.301	Prob(JB):	4.58e-307
Kurtosis:	4.236	Cond. No.	13.2

Figure 27. OLS Regression results 2.

	sum_sq	df	F	PR(>F)
C(type)	1105.029735	1.0	11637.792996	0.000000e+00
C(year)	101.336738	3.0	355.747889	1.993786e-224
C(type):C(year)	20.665719	3.0	72.548080	1.242532e-46
Residual	1699.922602	17903.0	NaN	NaN

Figure 28. ANOVA two ways table.

In Figure 28, the p-value is small for all variables and interactions. Therefore, we can reject the null hypothesis.

3. Data Preparation

It is our objective in this section to elaborate on all of the data preparation phases. The first step is to justify the approach we would take in the case of missing values. If there were missing values for average price we would evaluate by sorting into region, type, month and year to calculate its mean values. In a situation where all these features are being used, it is necessary to have at least one categorical variable because they are determinant in predicting our price as Chi-square test showed us the test of independence. Reversed engineering could be applied to the categorical variables such as region or type and once we have the mean price sorted by month, year and at least one categorical variable of greater precision, then we will be able to use the mode to tell us what categorical variables appear most in our data. In a scenario where we don't have any categorical variable the alternative is to drop the null values.

Table 6 summarizes the Feature Engineering:

Feature Engineering for numerical variables			
New data	**New Data type**	**From**	**Original Data type**
Month	integer	Date	object
Seasons	object	Month	integer
Day	date/time	Date	object

Table 5. Feature Engineering for numerical variables.

Figure 29 shows us the variation of average price throughout months:

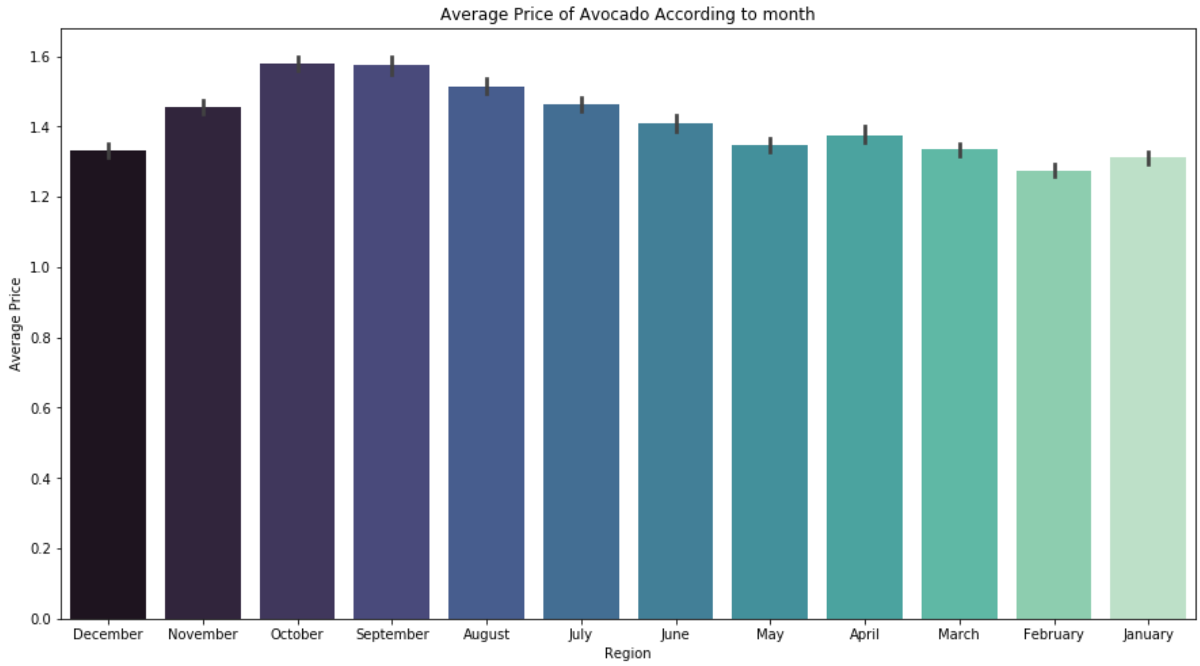

Figure 29. Average price by month.

Figure 30 shows us no sig. difference between the months of each particular season.

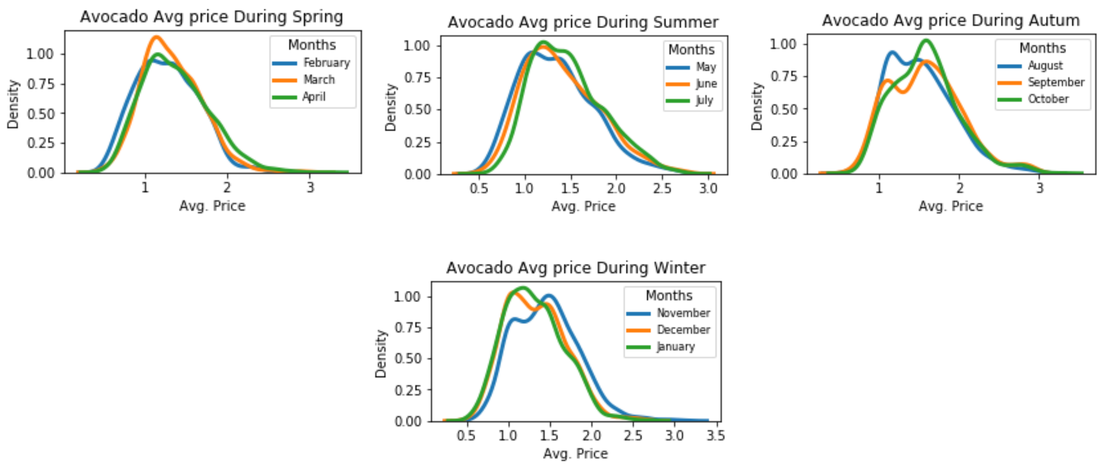

Figure 30. Multiple Line Charts comparison of seasonality.

The distribution of the seasons is shown in Figure 31.

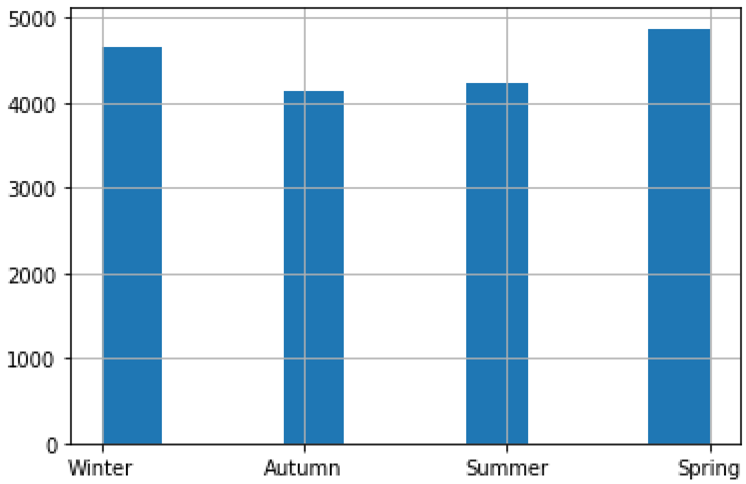

Figure 31. Multiple Line Charts comparison of seasonality.

Table 6 below summarizes our Feature Engineering for categorical variables:

Feature Engineering for Categorical variables	
Variable	**Method**
Region	Dummy-encoded
Type	Dummy-encoded
Season	Dummy-encoded
Year	Dummy-encoded

Table 6. Feature Engineering for categorical variables.

In order to avoid or minimize multicollinearity we used the heatmap correlation matrix to observe the correlation within the dependent variables as shown Figure 32. Table 8 summarizes all our actions and reasons for dropping the columns.

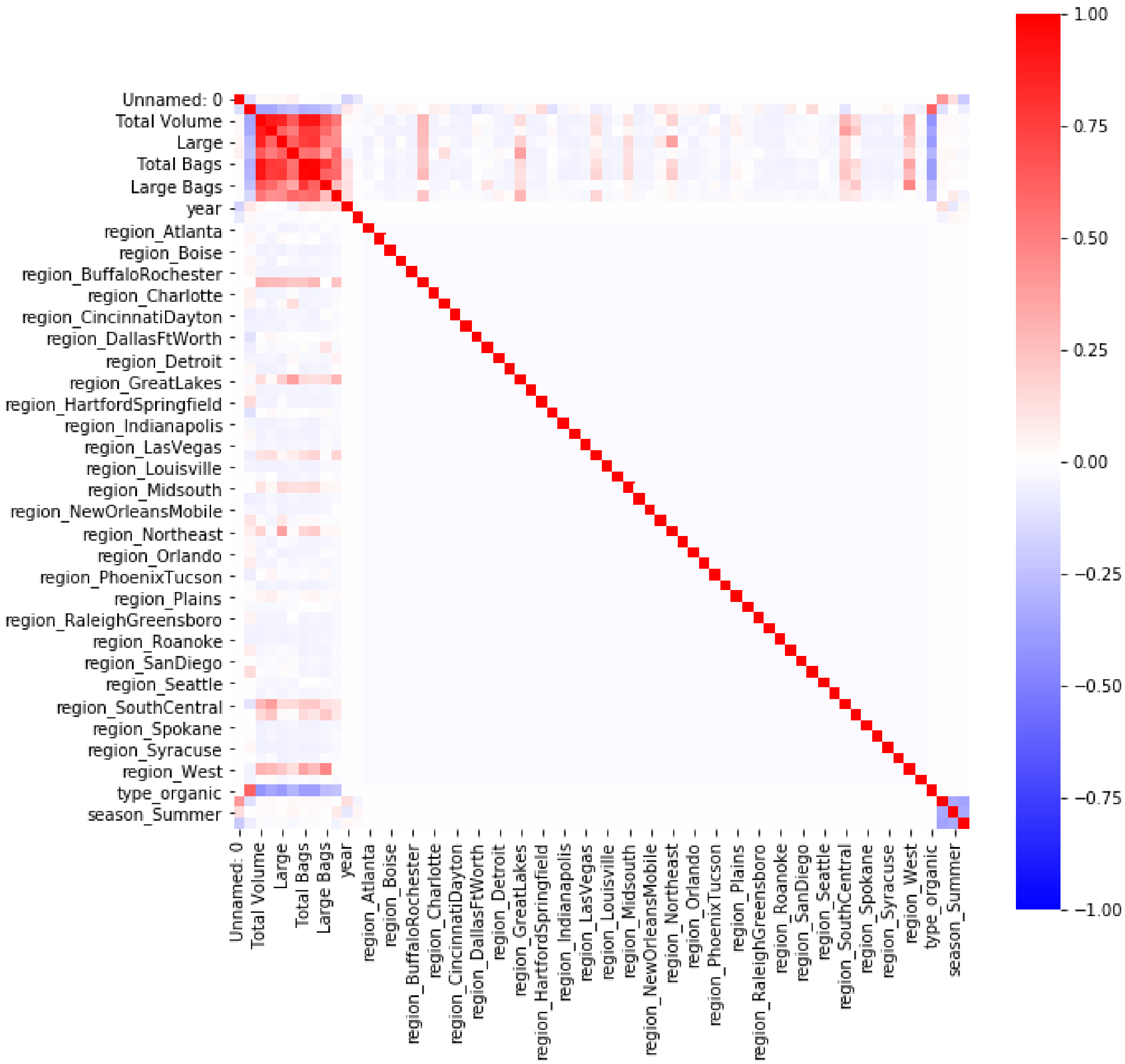

Figure 32. Heatmap correlation.

Reasons to drop the columns		
Data Item	**Data Type**	**Reason**
Total Volume	float	To avoid multicollinearity

Total Bags	float	To avoid multicollinearity
Small Bags	float	To avoid multicollinearity
Large Bags	float	To avoid multicollinearity
XLarge Bags	float	To avoid multicollinearity
Date	object	Seasons instead
Day	integer	Contributes nothing to the analysis
Unnamed: 0	integer	Contributes nothing to the analysis

Table 7. Reasons to drop the columns.

Figure 33 tells us that we still have some zero values and we will need to filtered:

	count	mean	std	min	25%	50%	75%	max
AveragePrice	17911.0	1.407619	0.404253	0.44	1.10	1.37	1.670	3.25
Small_Medium	17911.0	183807.409290	515105.860647	0.00	819.66	7824.43	101488.815	5160896.68
Large	17911.0	188223.112232	451985.648442	0.00	2909.61	26701.99	131755.215	5402444.45
ExtraLarge	17911.0	14551.234381	48817.536762	0.00	0.00	164.23	5736.735	804558.25
year	17911.0	2016.147898	0.939939	2015.00	2015.00	2016.00	2017.000	2018.00
...	...	...	...	...	...	...	...	...
region_WestTexNewMexico	17911.0	0.018704	0.135480	0.00	0.00	0.00	0.000	1.00
type_organic	17911.0	0.499916	0.500014	0.00	0.00	0.00	1.000	1.00
season_Spring	17911.0	0.272235	0.445123	0.00	0.00	0.00	1.000	1.00
season_Summer	17911.0	0.236614	0.425015	0.00	0.00	0.00	0.000	1.00
season_Winter	17911.0	0.260343	0.438834	0.00	0.00	0.00	1.000	1.00

Figure 33. Central Tendency Measures final dataset.

4. Models and Evaluation

Before analysing the test results we will plot the actual versus predicted and the residual analysis for the train data as shown below.

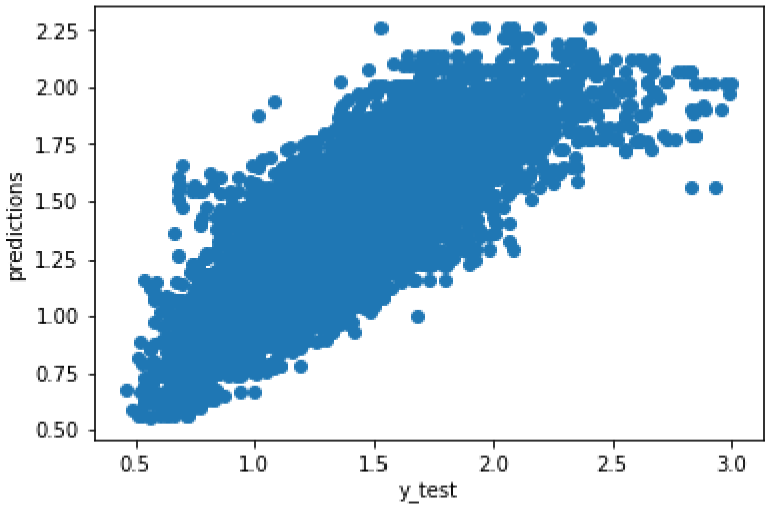

Figure 34. Scatterplot predictions for training data.

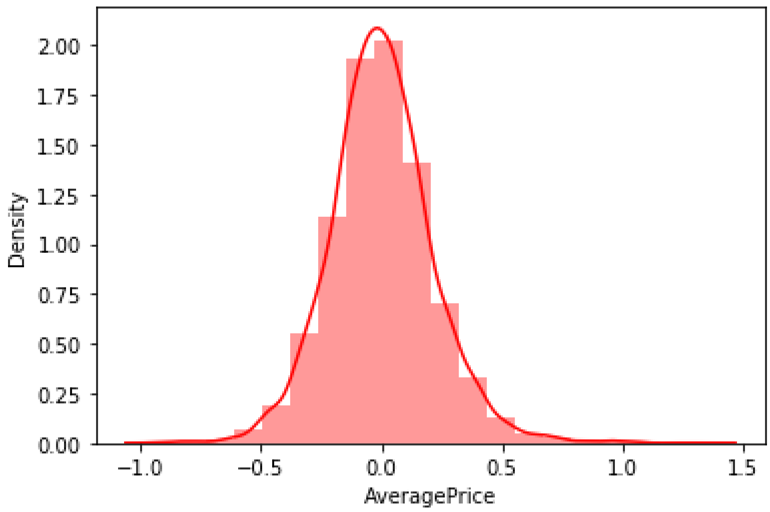

Figure 35. Residual analysis for training data.

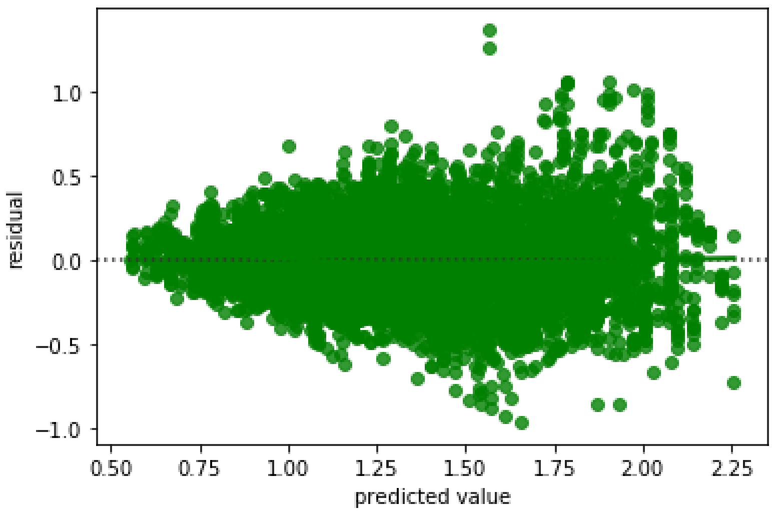

Figure 36. Scatterplot residuals actuals vs predicted.

In Figure 37, our model did not fit the expected line based on the train data. In Figure 38, we did not get the desirable straight red line. Based on these two conditions indicates the presence of heteroscedasticity.

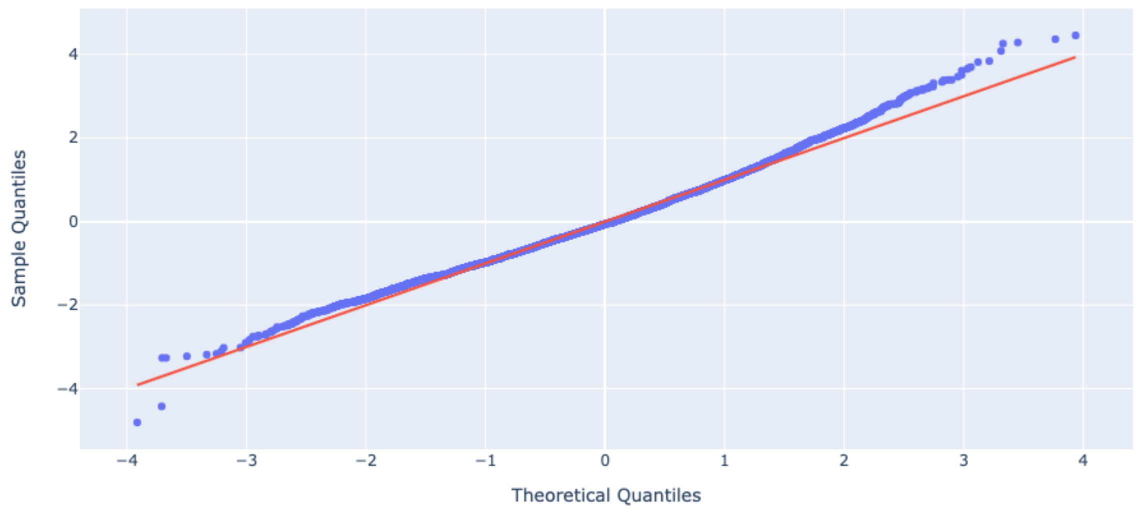

Figure 37. Theoretical Quantiles.

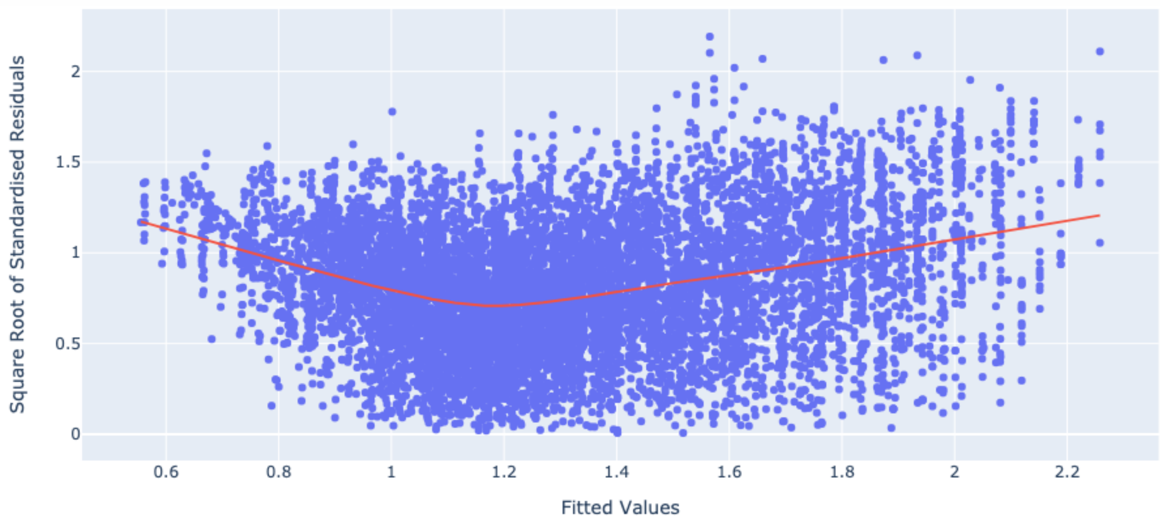

Figure 38. Scatterplot square root of standardised residuals.

We also tested the Durbin Watson test and we got a value of 1.9926 meaning that there is no autocorrelation in the residuals.

Figures 40 and 41 shows us that we should reduce its dimensionality to a level of 2 components.

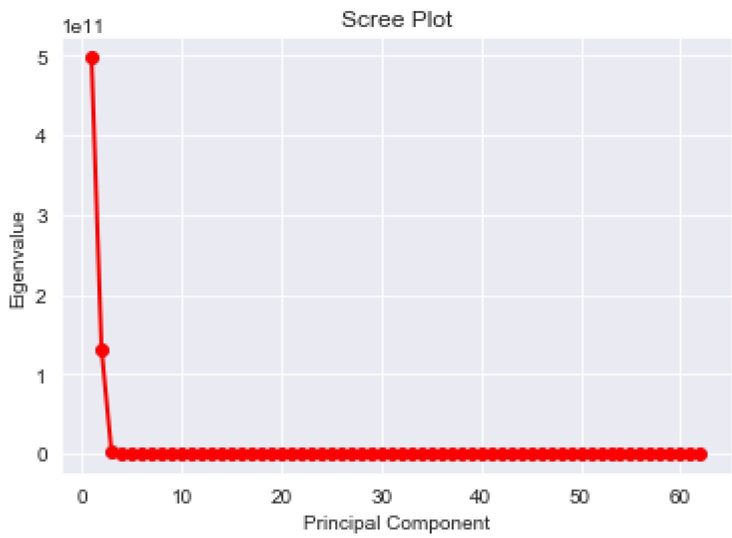

Figure 40. Scree Plot PCA analysis.

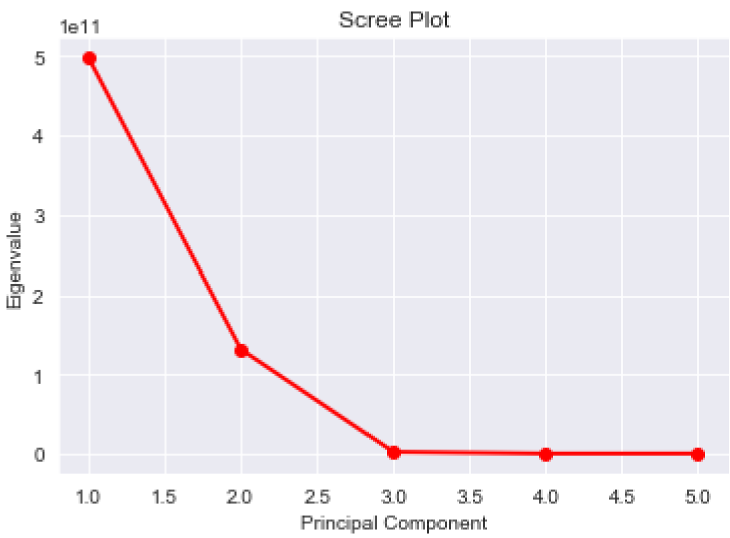

Figure 41. Scree Plot PCA analysis 2.

Table 8 is the summary of all models including PCA. It's clear that the dimensionality reduction approach PCA did not improve the analysis but actually decreased the models performance . One of the reasons we believe that is the amount of dummy variables that were made from the regions and all of them are relevant to the analysis. Not considering the dimensionality reduction, the three models performed quite similar. Lasso Regression with an alpha equal to 0.0001 performed slightly better than the others and as a result, we conclude that dimensionality reduction is not needed for this analysis and all the features included in this final version of the dataset are important to achieve a reasonable accuracy. It's probably that this dataset is not appropriate for Regression tasks as even with the Regression diagnostics we were not able to improve our model. It is out of the scope of this project but one proposed solution could be implementation of a Random Forest model which belongs to the category of ensemble methods, and can be used for classification and regression problems. Random Forest specifically is a powerful algorithm that divides the dataset into a subset of samples and generates multiple decision trees based on the mean prediction. Random Forest models are particularly adept at handling high levels of dimensionality in a dataset which will suit our dataset due to the magnitude of dummy encoded variables.

Regression models evaluation								
Models	**Split Type**	**Number of Features**	**Best Parameter (alpha)**	**Train Set Accuracy (R2)**	**Test Set Accuracy (R2)**	**Test Set MAE**	**Test Set MSE**	**Test Set RMSE**
Linear Regression	25% Simple	63	-	0.7048	0.6985	0.1595	0.0459	0.2144
Ridge Regression	25% Simple	63	1.0	0.7048	0.69855	0.1595	0.04596	0.2144
Lasso Regression	25% Simple	63	0.0001	0.7047	0.6989	0.1593	0.0459	0.2142

Linear Regression with PCA	25% Simple	2 + dependent variable	-	0.1140	0.1155	0.2881	0.1295	0.3599
Ridge Regression with PCA	25% Simple	2 + dependent variable	0.015	0.0855	0.0887	0.2905	0.1334	0.3653
Lasso Regression with PCA	25% Simple	2 + dependent variable	8.80E-06	0.114	0.1155	0.3073	0.1466	0.3829

Table 8. Regression models evaluation.

The conventional type has a lot more data than the organic. Therefore, it did not help our analysis as the model learned more about one than the other. The accuracy of the classification task was the same with or without the application of the alpha penalty. We believe one of the reasons for a high score is perhaps that the data is not well distributed and as a result can bias the model.

Logistic regression evaluation							
Model	**Conventional (0)**	**Organic (1)**	**Split Type**	**Number of Features**	**Best Parameter (alpha)**	**Train Set Accuracy (R2)**	**Test Set Accuracy (R2)**
Logistic Regression	8,956	3,458	25%	63	C': 100, 'penalty': 'l2'	1.000	0.9999

Table 9. Logistic regression evaluation.

In Figure 42, the confusion matrix confirms the accuracy of the model. It's incorrectly predicted at a total of four times.

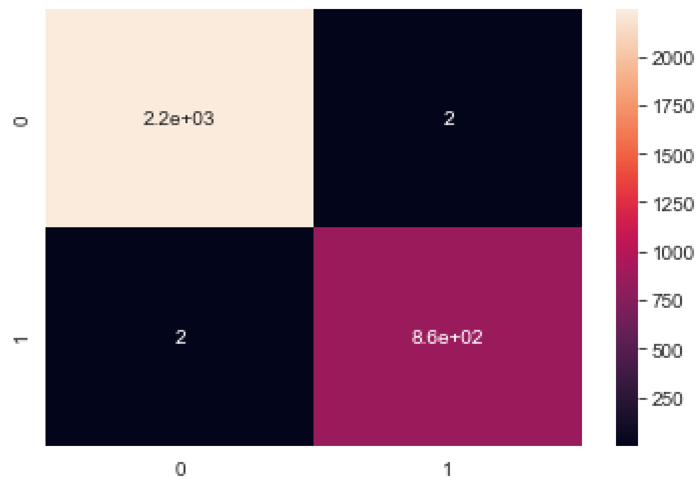

Figure 42. Confusion matrix.

5. References

Kiggins, J. (2018). *Avocado Prices*. Available at: https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho (Accessed 10 December 2020).

Gunnars, K. (2018). *12 Proven Health Benefits of Avocado.* Available at: https://www.healthline.com/nutrition/12-proven-benefits-of-avocado (Accessed 11 December 2020).

AGMRC. (2018). *Avocados.* Available at: https://www.agmrc.org/commodities-products/fruits/avocados (Accessed 12 December 2020).

Htoon, K. S. (2020). Log Transformation: Purpose and Interpretation. Available at: https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9 (Accessed 14 December 2020)

Chen, B. (2018). *What is One-Hot Encoding and how to use Pandas get_dummies function*. Available at: https://towardsdatascience.com/what-is-one-hot-encoding-and-how-to-use-pandas-get-dummies-function-922eb9bd4970 (Accessed 14 December 2020).

Ford, C. (2015). *Understanding Q-Q plots*. Available at: https://data.library.virginia.edu/understanding-q-q-plots/ (Accessed 14 December 2020)

Guest Blog. (2020). *Introduction to ANOVA for Statistics and Data Science (with COVID-19 Case Study using Python).* Available at: https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/ (Accessed 20 December 2020)

Garg, R. (2018). A Primer to Ensemble Learning - Bagging and Boosting. Available at: https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/#:~:text=Bagging%20is%20a%20way%20to,based%20on%20the%20last%20classification (Accessed 22 December 2020).

Kenton, W. (2019). *Durbin Watson Statistic Definition.* Available at: https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic (Accessed 27 December 2020).

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR, 12*, pp. 2825-2830.

Ruginski, I. (2016). Checking the assumptions of linear regression. *Accessed, 11*, p. 2018.

6. Appendix

6.1 Roles and Responsibilities

Roles and Responsibilities	
Task	**Task Owner**
Find and choose suitable dataset	Giuliano
EDA, Data Cleaning and Data Preparation	Eric, Giuliano
Introduction, Objectives and Statistical Tests Research	Conor and Hasan
Chi-square and ANOVA test	Giuliano, Eric
Statistical test for Regression	Giuliano and Hasan
Build Linear Regression Model	Conor and Giuliano
Build Ridge and Lasso Regression	Eric and Hasan
PCA	Eric and Giuliano
Build Logistic Regression	Conor, Hasan
ML Model Evaluation	Conor, Eric, Giuliano and Hasan
Conclusion	Conor, Eric, Giuliano and

	Hasan
Lead Python Programmer/Model Builder	Giuliano and Eric
ML Model Quality Analysis	Conor and Hasan

Table 10. Roles and responsibilities.