Analysis of Hass Avocado Price in California

Instructor: Daniel D. Gutierrez
Class: Spring 2020 – Introduction to Data Science
Chuyin Zhu

Table of Contents

1. Introduction

1.1 Background Introduction

Avocados! Not only are they crowned as one of the main agricultural crops (in terms of value) in California, but also they are beloved by millions of Californians and many more people around the world. As much of the attention has been drawn to its nutritional benefits and tastes, this report presents an analysis of another side about the Hass avocados, one of the most popular varieties. What are the factors that could potentially affect its price in California?

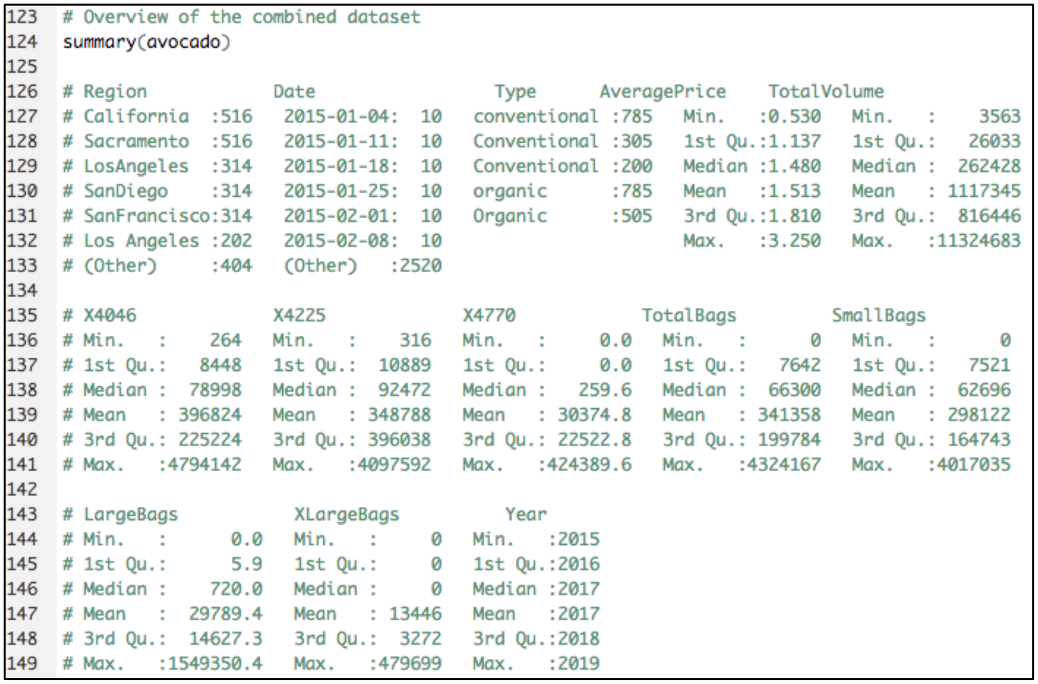
Avocados can be easily found in most groceries stores and in almost all year around. However, behind the price tags shown in front of the consumers' eyes, there is a story of how the price can be affected.

1.2 Dataset Preparation & Overview

The dataset of the report comes from two sources—Kaggle (Kaggle.com/neuromusic/avocado-prices) and Hass Avocado Board (hassavocadoboard.com). The first source provides data from 2015 to 2017, and the second one has data available from 2018 and 2019.

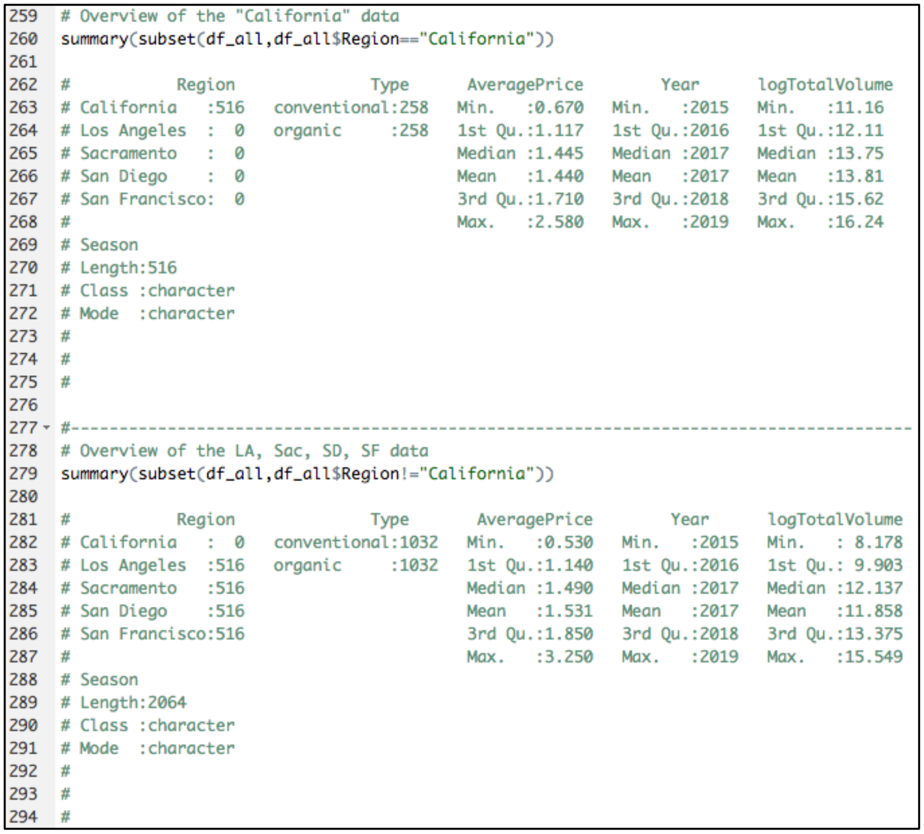
After all datasets of different years are reorganized and combined into one, a few variables are dropped, and some new predictors are created. The following page shows the "Original Dataset Summary (of the dataset that is not completely cleaned up and reorganized)" as well as the "Final Dataset Summary (of the dataset that is ready for analysis)". First, the "Season" predictor is created based on the "Date" variable in the original dataset. Second, variables of Hass avocados sold in different PLU sizes ("X4046 (small)", "X4225 (large)", and "X4770 (all size)") or in different bag sizes ("TotalBags", sum of "SmallBags", "LargeBags", and "XLargeBags") are dropped due to multicollinearity with "TotalVolume (lbs.)". "TotalVolume" equals to avocados sold in different PLU and bag sizes. Also, the sales volume of California in this dataset equals to the combined sales of four cities—Los Angeles, San Francisco, Sacramento, and San Diego. After (natural) log transformation, the new predictor "logTotalVolume" replaces the "TotalVolume" in the final dataset, representing the sales growth. In the "Final Dataset Summary", there are two types of avocados—conventional and organic. The "AveragePrice (US $)" is the price of a single Hass avocado. In total, from the year of 2015 to 2019, there are 2,580 observations, where 516 belong to California, and 2,064 to the four cities.

(Original Dataset Summary)



```
# Overview of the combined dataset
summary(avocado)

# Region             Date                Type         AveragePrice    TotalVolume
# California  :516   2015-01-04:  10   conventional :785   Min.   :0.530   Min.   :    3563
# Sacramento  :516   2015-01-11:  10   Conventional :305   1st Qu.:1.137   1st Qu.:   26033
# LosAngeles  :314   2015-01-18:  10   Conventional :200   Median :1.480   Median :  262428
# SanDiego    :314   2015-01-25:  10   organic      :785   Mean   :1.513   Mean   : 1117345
# SanFrancisco:314   2015-02-01:  10   Organic      :505   3rd Qu.:1.810   3rd Qu.:  816446
# Los Angeles :202   2015-02-08:  10                       Max.   :3.250   Max.   :11324683
# (Other)     :404   (Other)   :2520

# X4046              X4225              X4770              TotalBags          SmallBags
# Min.   :    264   Min.   :    316   Min.   :     0.0   Min.   :      0   Min.   :      0
# 1st Qu.:   8448   1st Qu.:  10889   1st Qu.:     0.0   1st Qu.:   7642   1st Qu.:   7521
# Median :  78998   Median :  92472   Median :   259.6   Median :  66300   Median :  62696
# Mean   : 396824   Mean   : 348788   Mean   : 30374.8   Mean   : 341358   Mean   : 298122
# 3rd Qu.: 225224   3rd Qu.: 396038   3rd Qu.: 22522.8   3rd Qu.: 199784   3rd Qu.: 164743
# Max.   :4794142   Max.   :4097592   Max.   :424389.6   Max.   :4324167   Max.   :4017035

# LargeBags            XLargeBags          Year
# Min.   :      0.0   Min.   :     0   Min.   :2015
# 1st Qu.:      5.9   1st Qu.:     0   1st Qu.:2016
# Median :    720.0   Median :     0   Median :2017
# Mean   :  29789.4   Mean   : 13446   Mean   :2017
# 3rd Qu.:  14627.3   3rd Qu.:  3272   3rd Qu.:2018
# Max.   :1549350.4   Max.   :479699   Max.   :2019
```

(Final Dataset Summary: California, Los Angeles, Sacramento, San Diego, San Francisco)

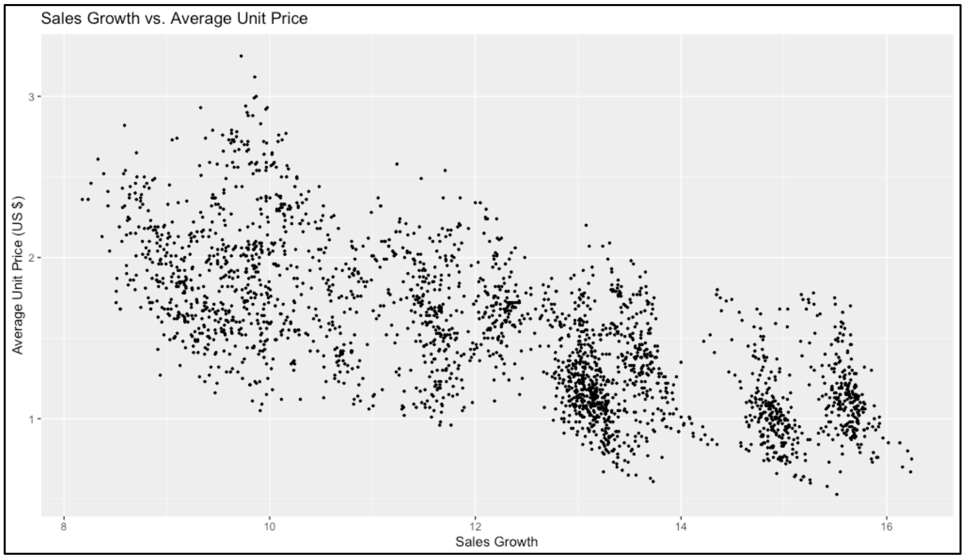


```
# Overview of the "California" data
summary(subset(df_all,df_all$Region=="California"))

#            Region              Type         AveragePrice       Year        logTotalVolume
# California    :516   conventional:258   Min.   :0.670   Min.   :2015   Min.   :11.16
# Los Angeles   :  0   organic     :258   1st Qu.:1.117   1st Qu.:2016   1st Qu.:12.11
# Sacramento    :  0                      Median :1.445   Median :2017   Median :13.75
# San Diego     :  0                      Mean   :1.440   Mean   :2017   Mean   :13.81
# San Francisco:  0                       3rd Qu.:1.710   3rd Qu.:2018   3rd Qu.:15.62
#                                         Max.   :2.580   Max.   :2019   Max.   :16.24
# Season
# Length:516
# Class :character
# Mode  :character
#
#
#

#-------------------------------------------------------------------------------------
# Overview of the LA, Sac, SD, SF data
summary(subset(df_all,df_all$Region!="California"))

#            Region              Type          AveragePrice       Year        logTotalVolume
# California    :  0   conventional:1032   Min.   :0.530   Min.   :2015   Min.   : 8.178
# Los Angeles   :516   organic     :1032   1st Qu.:1.140   1st Qu.:2016   1st Qu.: 9.903
# Sacramento    :516                       Median :1.490   Median :2017   Median :12.137
# San Diego     :516                       Mean   :1.531   Mean   :2017   Mean   :11.858
# San Francisco:516                        3rd Qu.:1.850   3rd Qu.:2018   3rd Qu.:13.375
#                                          Max.   :3.250   Max.   :2019   Max.   :15.549
# Season
# Length:2064
# Class :character
# Mode  :character
#
#
#
```

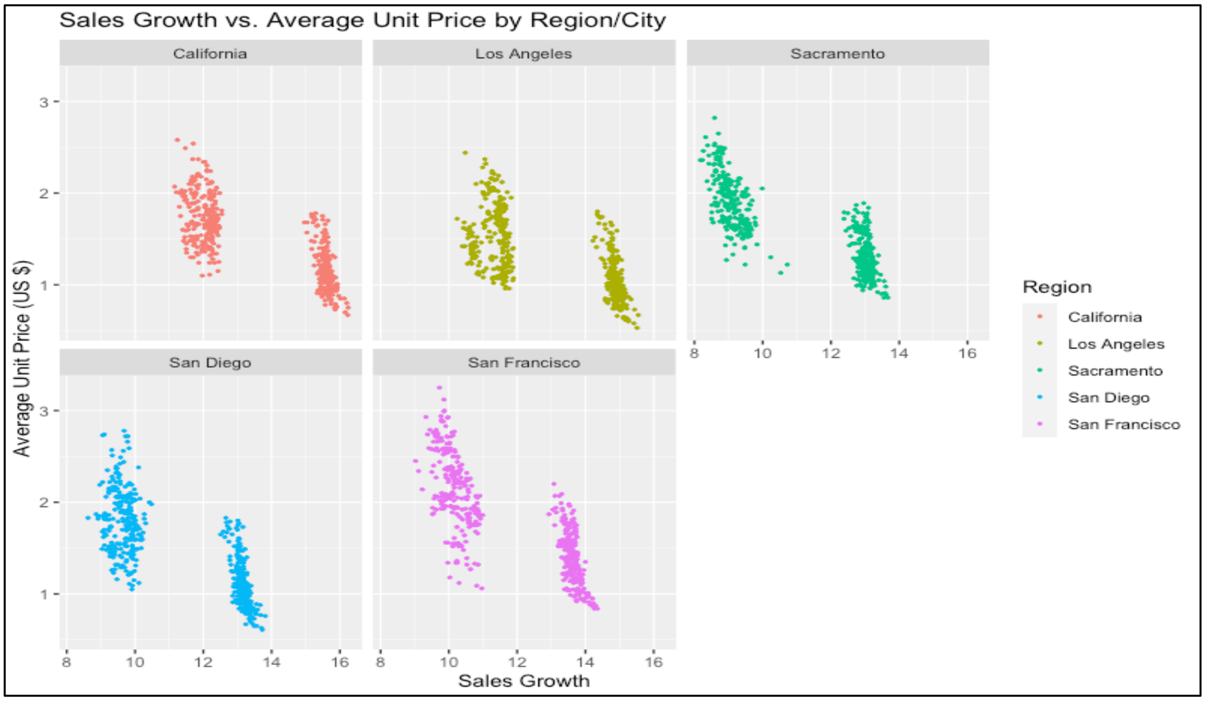
2. Exploratory Data Analysis

2.1 Sales Growth

The scatter plot “Sales Growth vs. Average Unit Price” below observes a pattern trend between average unit price and sales growth of all the observations. As the sales growth increases, the average price of a single avocado is likely to decrease. The points are less spread.



In the following scatterplot, such a downward-sloping pattern also reflects at the California and city level. However, the pattern in each plot is broken into two parts, which indicates that a group of avocados has

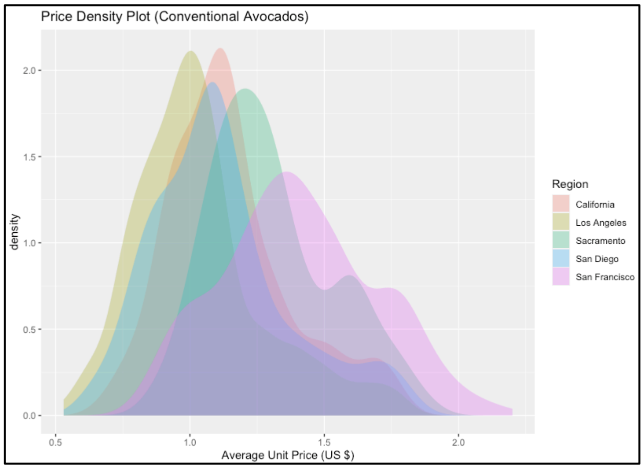


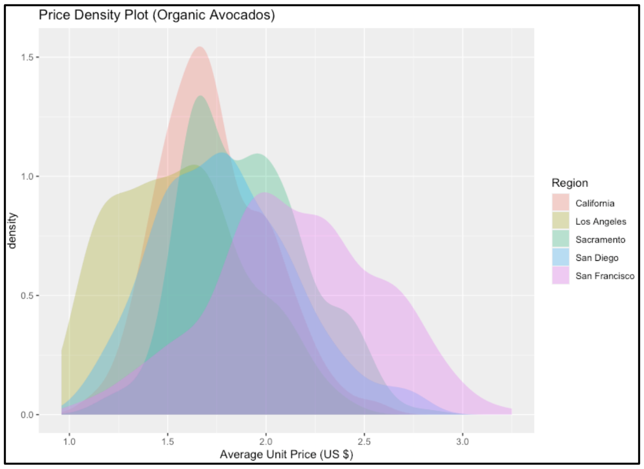
a higher sales growth at a lower price range, mainly driven by Los Angeles. Another group has a relatively lower sales growth and is more expensive, favored by consumers in Sacramento and San Diego. Overall, Hass avocados are most expensive in San Francisco, compared to the other three cities. Sales growth and average unit price vary across cities.

2.2 Conventional vs. Organic

Are organic Hass avocados always preferred over the conventional ones? Overall, Conventional avocados are more affordable than the organically grown ones. However, a lot of avocados consumers may consider the latter ones to have (much) fewer pesticides and thus are willing to spend more money on those marked as organic. While others seem to not show much concerns on whether the avocados are grown organically or conventionally.

The plot on the left side below shows the price density of the conventional Hass avocados in different region/cities. The price range of a single conventional Hass is from around $0.50 to $2.40, with Los Angeles on the cheaper end of the average unit price range and San Francisco on the more expensive end. Most of the conventional Hass avocados sold in Los Angeles are $1.00 each, followed by San Diego, Sacramento, and San Francisco. Conventional avocados in San Francisco are sold at a unit price from as low as around $0.60 to $2.40. On average, as the majority conventional avocados are sold at from $0.75 to $1.50 in California, it suggests that Los Angeles and San Diego have big markets for the conventional.

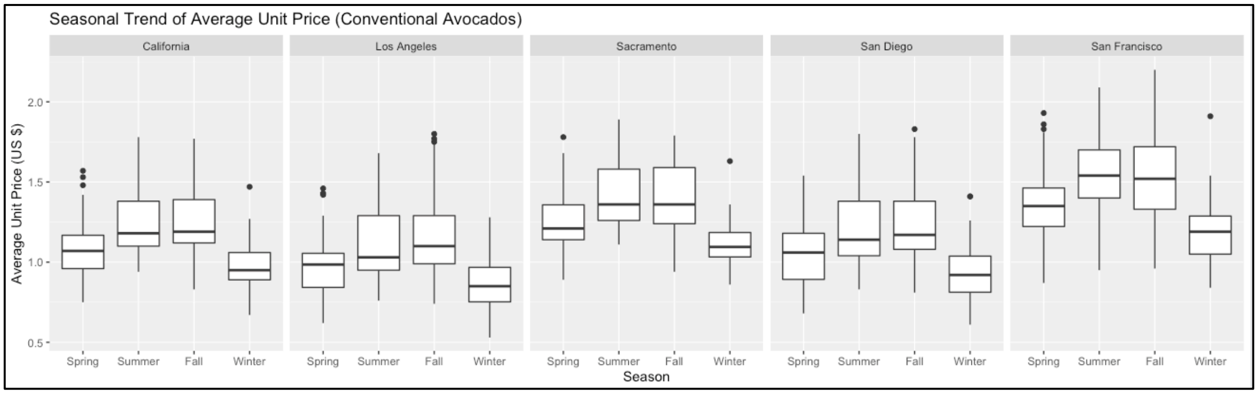


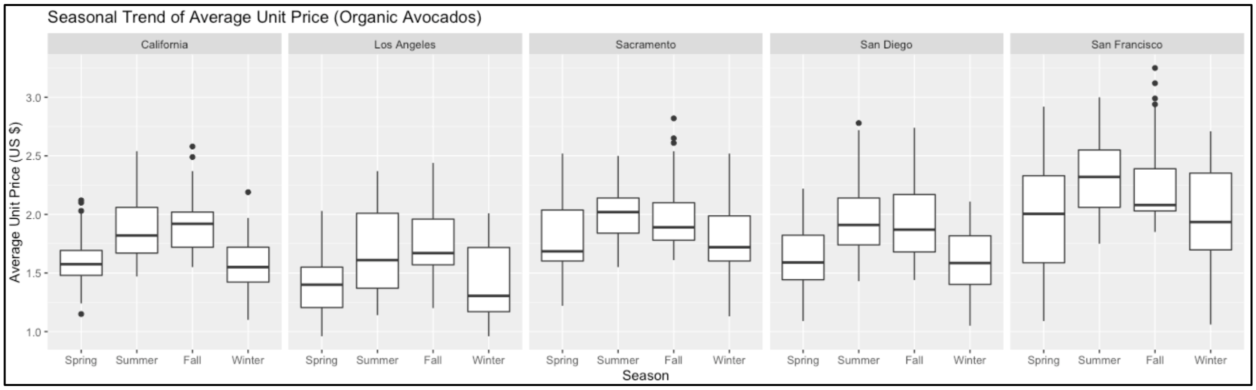


The plot on the right side above shows the price density for the organic avocados. In San Francisco, a single organic Hass avocado can sometimes be sold as high as $3.00 or more. Although people can still buy the organic ones starting at $1.00 (or even less) each, the organic ones there are likely to be more expensive than the other three cities. Organic Hass avocados in Sacramento have two “peak” average unit prices—approximately at $1.60 and $2.00. Not only does Los Angeles have the cheapest conventional Hass avocados, but also its organic ones are most affordable. The average unit price of the organic avocados has a wider range than the conventional ones. Whether organic or conventional, Hass avocados appear to be more expensive in northern cities than in the southern ones in California

2.3 Seasonality

Previous exploratory data analysis has demonstrated that Hass avocados' unit price fluctuates in sales growth, types (conventional or organic), and cities in California. How about in different seasons?





In California, spring lasts from March to May, summer from June to August, fall from September to November, and winter from December to February in the following year. For conventional Hass avocados, they are most affordable in winter, followed by spring. Their average unit price is above $1.00 during summer and fall in California overall. Among the four cities, conventional Hass avocados in Los Angeles seem to be the cheapest in all of the four seasons, and the ones in San Francisco are the most expensive all year round. The average unit price in summer and fall is likely to have a larger fluctuation than in the other two seasons.

On the other hand, the average unit price of the organic Hass avocados shows a similar seasonality pattern. On average, the unit price fluctuation is within $0.50 for the conventional ones in each season. However, for the organic Hass, there are times when the unit price fluctuation is more than $0.50, for example, summer and winter time in Los Angeles, fall in San Diego, as well as spring and winter in San Francisco. Price outliers can be spotted more often during spring and winter for the conventional avocados. Meanwhile, outlier cases happen more frequently in the fall for the organic ones.

Seasonality influences the average unit price of both the conventional and organic avocados. Although in each of the region/cities, Hass avocados are less affordable in summer and fall than in spring and winter, the average unit price demonstrates different seasonal fluctuation patterns, depending on if the avocados are conventional or organic, and also on where (or which city) they are sold.

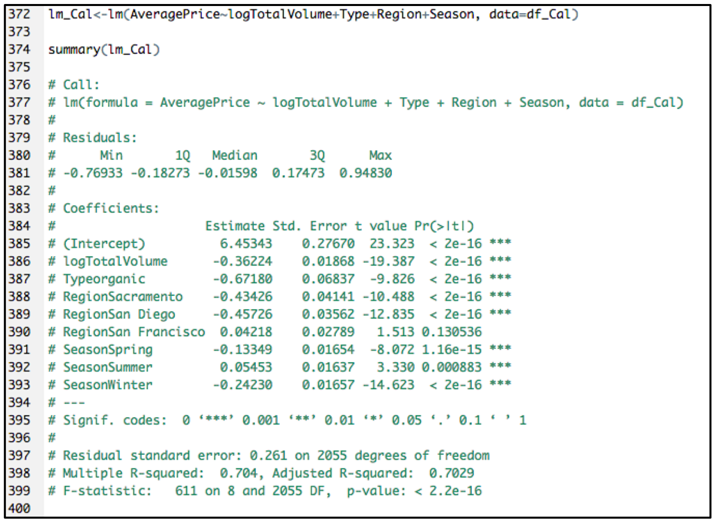
3. Machine Learning Algorithm

The report will use two multi-linear regression (MLR) models to predict and discover what factors can make significantly impact on the average unit price of avocados, with existing predictors.

3.1 Model I, Diagnosis, and Prediction

In the first MLR model (see result below), a subset of the dataset is used to explore whether the sales growth, types sold, cities where the Hass avocados are sold, and seasons when they are sold have any contribution to the variation of the average unit price.

(Model I: Average Unit price ~ log (Total Volume) + Type + Region + Season)

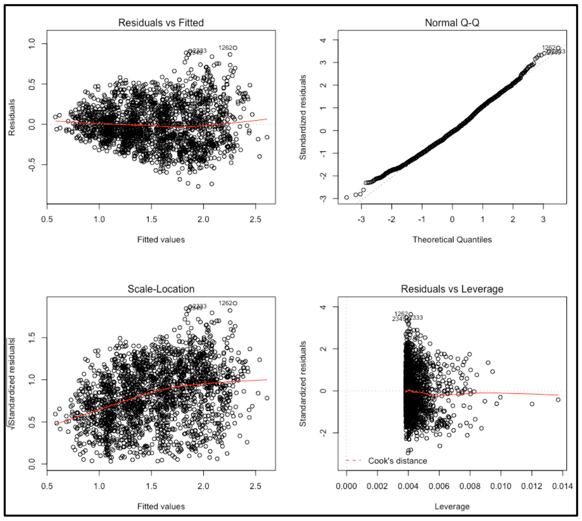


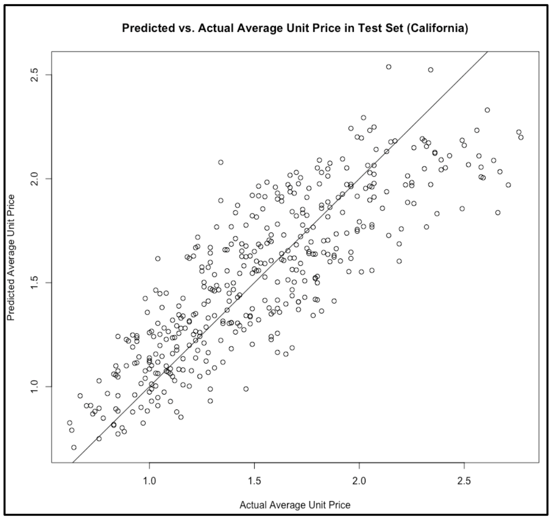
```
lm_Cal<-lm(AveragePrice~logTotalVolume+Type+Region+Season, data=df_Cal)

summary(lm_Cal)

# Call:
# lm(formula = AveragePrice ~ logTotalVolume + Type + Region + Season, data = df_Cal)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.76933 -0.18273 -0.01598  0.17473  0.94830
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)          6.45343    0.27670  23.323  < 2e-16 ***
# logTotalVolume      -0.36224    0.01868 -19.387  < 2e-16 ***
# Typeorganic         -0.67180    0.06837  -9.826  < 2e-16 ***
# RegionSacramento    -0.43426    0.04141 -10.488  < 2e-16 ***
# RegionSan Diego     -0.45726    0.03562 -12.835  < 2e-16 ***
# RegionSan Francisco  0.04218    0.02789   1.513 0.130536
# SeasonSpring        -0.13349    0.01654  -8.072 1.16e-15 ***
# SeasonSummer         0.05453    0.01637   3.330 0.000883 ***
# SeasonWinter        -0.24230    0.01657 -14.623  < 2e-16 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.261 on 2055 degrees of freedom
# Multiple R-squared:  0.704, Adjusted R-squared:  0.7029
# F-statistic:   611 on 8 and 2055 DF,  p-value: < 2.2e-16

```

In this subset, a training set of 1,651 (80% of the total 2,064 observations in the subset) is used to build model for prediction. In Model I, the regression has an adjusted R-squared of 0.7029, indicating that approximately 70% of the average unit price of the Hass avocados can be predicted or explained by the predictors used. Based on the 95% significance, sales growth ("logTotalVolume"), type ("Typeorganic"), city ("RegionSacramento" and "RegionSan Diego"), and season ("SeaonSpring", "SeasonSummer", "SeansonWinter") are all statistically significant and thus playing an important role on affecting the average unit price in California. Other predictors—"Typeconventional", RegionLos Angeles", SeasonFall"—are dummies and thus automatically dropped in the MLR analysis. However, their impact is collectively reflected through the interception, which is also statistically significant.

Overall, sales growth is negatively correlated with average unit price. For instance, for every percentage increase in sales growth, average unit price is likely to reduce by $0.36. Organic Hass avocados have negative impact on average unit price. The Sacramento and San Diego markets help drive the average unit price down, while the San Francisco market leads no significant impact here. Hass avocados in spring and winter are more affordable.

Below on the left side are the set of diagnostic plots of Model I. Residuals are randomly spread along the fitted values in both "Residuals vs. Fitted" and "Scale-Location" plots. They are also normally distributed in the "Normal Q-Q" plot. Outliers are within the Cook's distance and thus not influential.



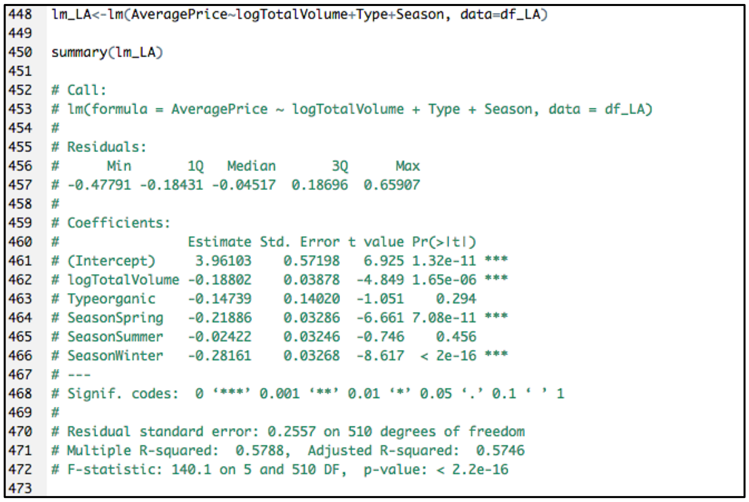


The trained Model I is then used to make prediction on the test set (20% of the total 2,064 observations in the subset). The plot on the right side above shows the predicted response values against the actual ones in the test set. The correlation between these two sets is high (r = 0.8389). In all, Model I is a good model in predicting the average unit price of Hass avocados in California.

3.2 Model II, Diagnosis, and Prediction

Model II analyzes only the observations in Los Angeles from the year of 2015 to 2019. By dropping the predictor "Region", a trained model built based on sales growth, type, and seasonality factors.

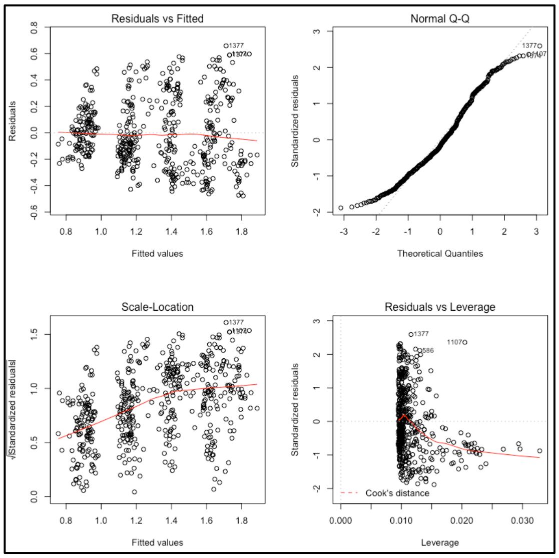
(Model II: Average Unit price ~ log (Total Volume) + Type + Season)

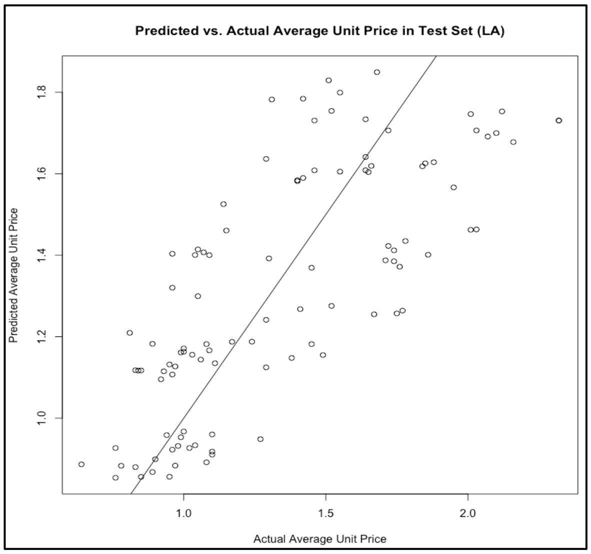


```
lm_LA<-lm(AveragePrice~logTotalVolume+Type+Season, data=df_LA)

summary(lm_LA)

# Call:
# lm(formula = AveragePrice ~ logTotalVolume + Type + Season, data = df_LA)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.47791 -0.18431 -0.04517  0.18696  0.65907
#
# Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
# (Intercept)      3.96103    0.57198   6.925 1.32e-11 ***
# logTotalVolume -0.18802    0.03878  -4.849 1.65e-06 ***
# Typeorganic    -0.14739    0.14020  -1.051    0.294
# SeasonSpring   -0.21886    0.03286  -6.661 7.08e-11 ***
# SeasonSummer   -0.02422    0.03246  -0.746    0.456
# SeasonWinter   -0.28161    0.03268  -8.617  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.2557 on 510 degrees of freedom
# Multiple R-squared:  0.5788,  Adjusted R-squared:  0.5746
# F-statistic: 140.1 on 5 and 510 DF,  p-value: < 2.2e-16
```

Model II has a much lower adjusted R-squared (0.5746), compared to Model I. Only about 57% of the average unit price of avocados in Los Angeles can be predicted in this model. Similar to Model I, sales growth still influences the average unit price. For every percentage of reduction in sales growth, the average unit price is likely to increase by $ 0.19. Organic avocados as well have negative but less impact on the change of average unit price. The avocado market during spring and winter is more competitive, driving down the average unit price; however, summer plays no significant part on affecting it.

The diagnosis on the left side below shows that the residuals in the "Residuals vs. Fitted" plot





randomly spread along the horizontal red line, meaning that, in Model II, the residuals capture what the existing predictors have not been able to. However, residuals deviate on both ends of the upward sloping diagonal line in the "Normal Q-Q" plot. Also in the "Scale-Location" plot, residuals start to spread wider, resulting in a non-horizontal red line. Few residuals are identified as outliers in the "Residuals vs. Leverage" plot, yet they are not influential in Model II. Another correlation test is run between the predicted and actual average unit price in the Los Angeles test set. The result (0.7533) indicates a strong correlation, and thus Model II is a valid model to make prediction on the average unit price of the Hass avocados in Los Angeles.

4. Conclusion & Reflection

According to the two MLR models, in California, the price tagged on each single Hass avocado can be largely determined by in which city and during which season it is sold. It also depends on how fast the avocado consumption grows. Such a conclusion may also be applied specifically in Los Angeles. However, the avocado consumers in Los Angeles may not exactly be the same as those in the other three cities because of the variation in demographics, preferences, and purchasing behaviors on avocados. As a result, the California model (Model I) can somehow represent Los Angeles, but the Los Angeles model (Model II) is more locally specific.

Based on Model I and II, the same set of predictors that can significantly predict the overall average unit price of the Hass avocados in California may not necessarily have the same and strong prediction ability when the market is narrowed down to Los Angeles. The reasons can also due to the fact that the sample population in Model II (516) is much smaller than that in Model I (2,064). In addition, fewer predictors in Model II can potentially weaken the prediction ability. Furthermore, to increase the prediction ability of Model II, more locally specific predictors may be introduced into the model later. For example, other than the Hass avocados, are there other avocado brands available in the same market? Finally, other algorithms may also worth exploring in order to find a better prediction model and/or tell the story from different prospectives.