Northeastern University

Avocado Sales

Final Project Report

Authored By: Neha Thakur, Praparna Moharana, Prateek Singh

Course: ALY 6010 - Probability Theory & Introductory Statistics

Instructor’s Name: Prof. Amin Karimpour

Submission Date: December 18,2021

Table of Contents

Abstract

This assignment is associated with the course “Probability Theory and Introductory Statistics”. As a data analyst, it is important not only to process complex data in a meaningful way but also to summarize and visualize it effectively for better analysis. Languages like R can enable us to utilize our knowledge in statics to explore, process, visualize and analyze data. This assignment gives an opportunity to use R and derive statistical output for various hypothesis based on the dataset.

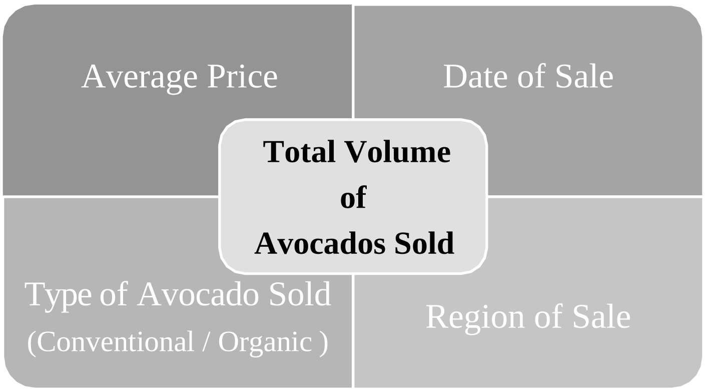
This report is final compilation of descriptive, analysis exploratory analysis, hypothesis testing, correlation and regression analysis conducted in the project.

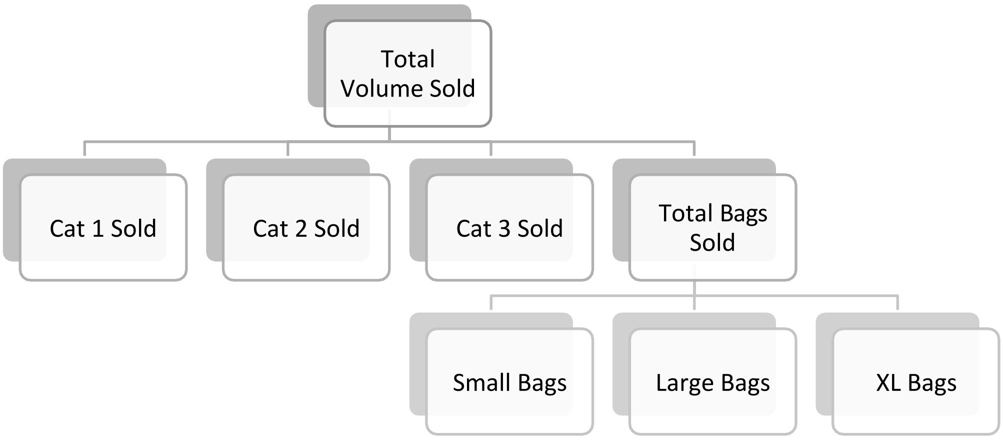
The dataset we have selected for this project is the “Avocado Sales” data. We found this dataset intriguing because it is relatable as we see Avocados being sold around us. We can easily understand and observe its implication. The model can be easily replicated to any retail case. This also gives an opportunity to predict future sales. These predictions can help the businesses in financial planning, inventory management and marketing management.

Avocado Sales

The avocado, is a tree likely originating from south-central Mexico, is classified as a member of the flowering plant family . The fruit of the plant, also called an avocado (or avocado pear or alligator pear), is botanically a large berry containing a single large seed. Avocados are cultivated in tropical and Mediterranean climates of many countries, with Mexico as the leading producer of avocados in 2019, supplying 32% of the world total. Depending on the variety, avocados have green, brown, purplish, or black skin when ripe, and may be pear-shaped, egg-shaped, or spherical. Commercially, the fruits are picked while immature, and ripened after harvesting.

In this project we are going to, look into various dimensions of avocado sales. We will analyze the trend of sales for three categories of Hass variety of avocados, which are either sold individually or in bags of small, large and extra-large size. This data has been picked up from Kaggle.com. We will analyze total volumes of avocados sold in different regions and also have a look into conventional and organic types of avocados.


Average Price
Date of Sale
Total Volume of Avocados Sold
Type of Avocado Sold (Conventional / Organic)
Region of Sale



This type of analysis can be replicated on different products in the retail industry and hence this dataset provides a huge scope of learning. This information can help Avocado companies to manage inventories, plan marketing campaigns and bring operational efficiencies in supply chain.

In the first part of the project, Milestone 1, we conducted exploratory data analysis. This gave us a good sense of the data. From the exploratory analysis, we have figured out some specific questions that we would want to ask of your data or the population as a whole. We will accomplish this using inferential statistics and hypothesis testing. In this assignment, Milestone 2, we will identify specific questions about the data or population parameters, conduct inferential testing on data and infer its results We will conduct one and two sample tests on price and volumes sold based on organic and conventional types.

Aim of Project

- To understand the trend of avocado sales over a period.
- To analyze differences in avocado sales depending on different categories and types.
- To identify high selling regions for avocado.
- To build a model for predicting future sales of avocado.

Data Set

The complete dataset includes 15207 observations with 13 variables mentioned below.

Variable	Data Type	Description
Date	Date	Date of each observation
Avg_Price	Numeric	Average of price of each avocado
Total_Volume	Numeric	Volume of Avocados sold in lbs
Cat1_Sales	Numeric	Avocados sold of type PLU 4046 in lbs
Cat2_Sales	Numeric	Avocados sold of type PLU 4225 in lbs
Cat3_Sales	Numeric	Avocados sold of type PLU 4770 in lbs
Total_Bags	Numeric	Total numbers of bag sold in lbs
Small_Bags	Numeric	Avocados sold in size bags in lbs
Large_Bags	Numeric	Avocados sold in large size bags in lbs
XLarge_Bags	Numeric	Avocados sold in extra-large size bags in lbs
Type	character	Types of Avocados [‘Conventional’, ’Organic’]
Year	Numeric	Years [2015,2016,2017,2018]
Region	character	Regions where Avocados are sold

Analysis

Summary of Variables

This is a summary table of all quantitative variables in the dataset and their 15207 observations. The minimum price for an avocado on a day of sales is $0.44 while maximum is

$3.25. On an average Cat1 and Cat2 avocados are sold more than Cat3 Avocados. Average sale of Cat1 avocado is 81360 lbs, it is 89186 lbs for Cat2 and its only 6869 lbs for Cat3. Small bags of avocados seem to be more popular compared to large and X large bags by their average sales. There are two types of avocados sold, Organic and Conventional.

Table 1.
Summary of Avg Price and Volume of Avocado Sales

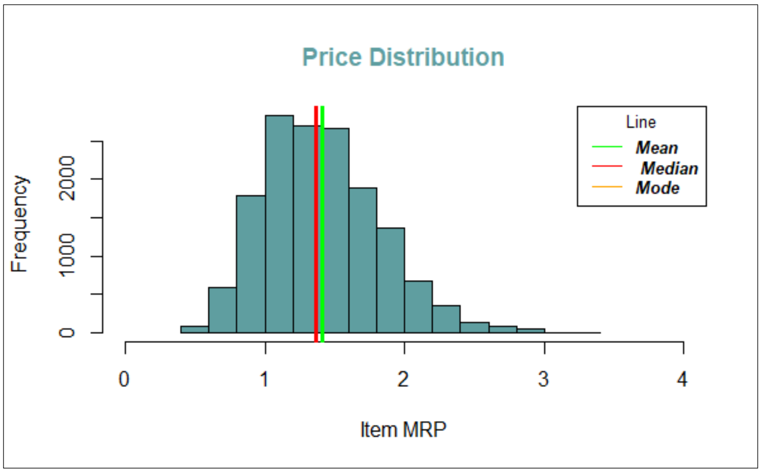
Variables	Mean	Sd	Min	Max	Range
Avg_Price (in $)	1.41	0.41	0.44	3.25	2.81
Total_Volume (in lbs)	249485	434375	85	5470227	5470143
Cat1_Sales (in lbs)	81360	198519	0	2914047	2914047
Cat2_Sales (in lbs)	89186	166241	0	2283465	2283465
Cat3_Sales (in lbs)	6869	19871	0	279630	279630
Total_Bags (in lbs)	72070	143734	0	2701610	2701610
Small_Bags (in lbs)	55090	124677	0	2656630	2656630
Large_Bags (in lbs)	16075	42268	0	706053	706053
XLarge_Bags (in lbs)	906	3875	0	61317	61317

Using these quantitative variables and some questions we will define hypothesis and test them statistically. The test will help us confirm if the observations from the population stated in the table below are same for any sample extracted from the data.

This report represents one sample and two sample t-test for variables “Volume Sold” and “Average Price” based on various categories. The following flowchart illustrates the steps performed in these tests.

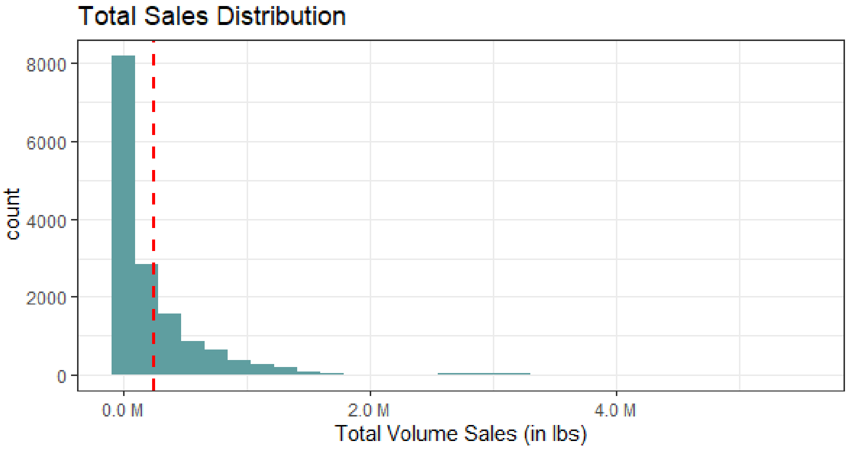
Exploratory Data Analysis

Exploratory data analysis helps to gain deeper understanding of dataset and extracting information from data as well identify trends as well as patterns which summarizes the main characteristics of data by visualization.



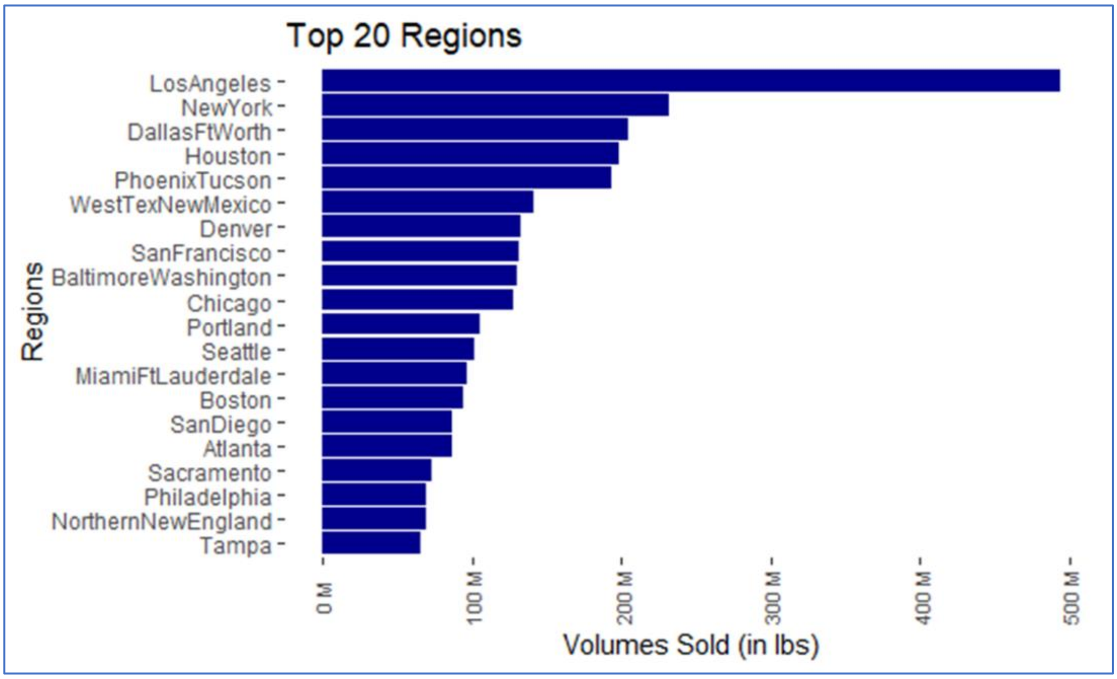
Plot 1: Histogram of Average Price Distribution

In above histogram plot shows the average price distribution of avocados from which we can infer that the average price range for an avocado, on a given day, starts from $0.4 to $3.2. The average price is $1.4 represented by the red line in the plot. The price distribution of avocados is almost a normal distribution. Mean, median and mode for the price almost overlap.



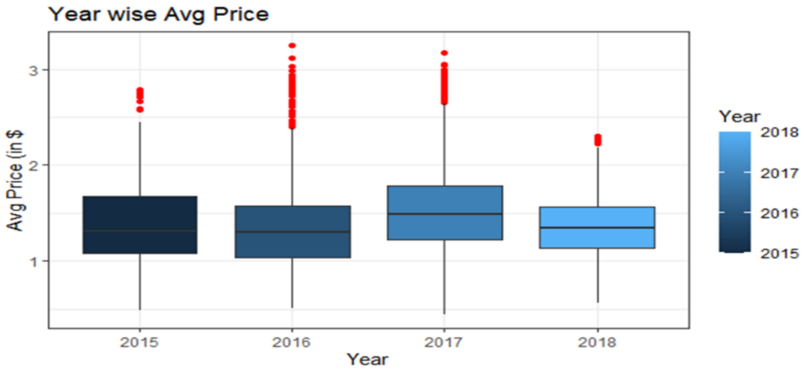
Plot 2: Histogram of Total Sales Distribution

The above histogram plot shows the distribution of total items sold from which we can infer that the mean of total volumes sold is around 200 thousand lbs. We can also observe some very high values. We will create a bar plot to look into these high sales volumes.



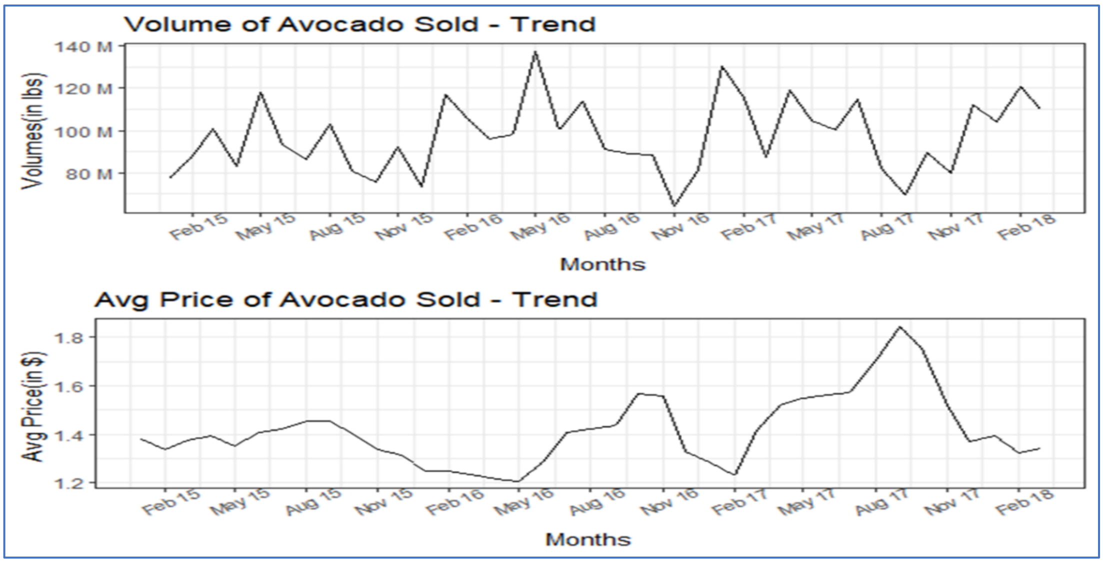
Plot 3: Bar plot of Top 20 Regions against Volume sold (in lbs.)

The above bar plot shows the volume of avocados sold in top 20 regions in descending order. Los Angeles has highest number of Avocados sold which has 500 million lbs. total sale of avocados which is much higher than other regions whereas Tampa has the least number of Avocados sold which we infer from the graph.



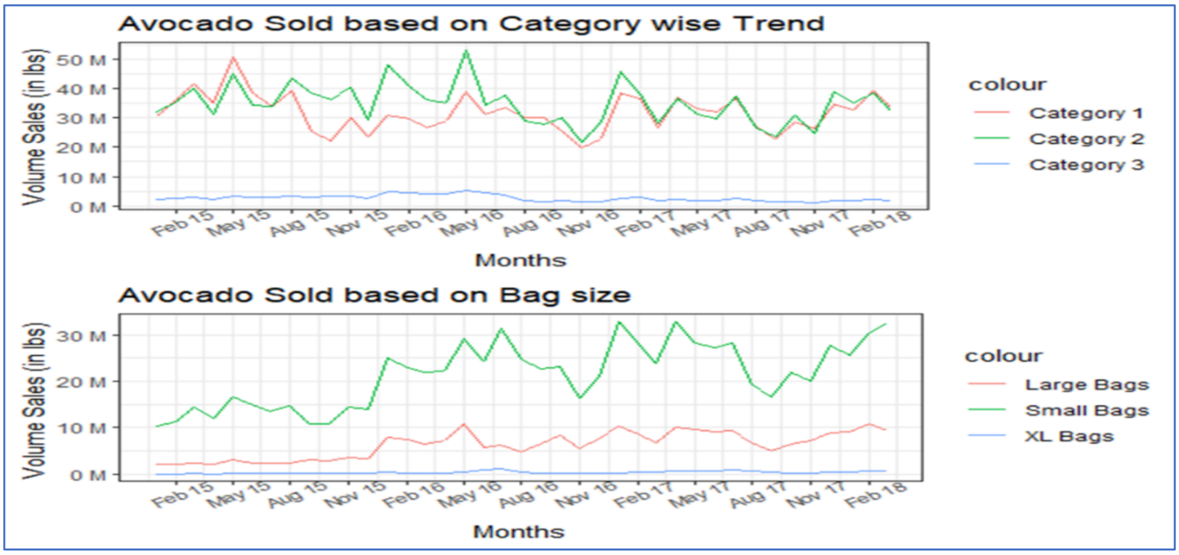
Plot 4: Box plot of Average price of Avocadoes each year

The boxplot depicts there is not much of a difference in average price of avocados each year which is almost similar in 2015,2016 and 2018. It is observed that 2017 a slight increase in average price of avocadoes whereas 2016 there is a slight decrease in average price of avocadoes



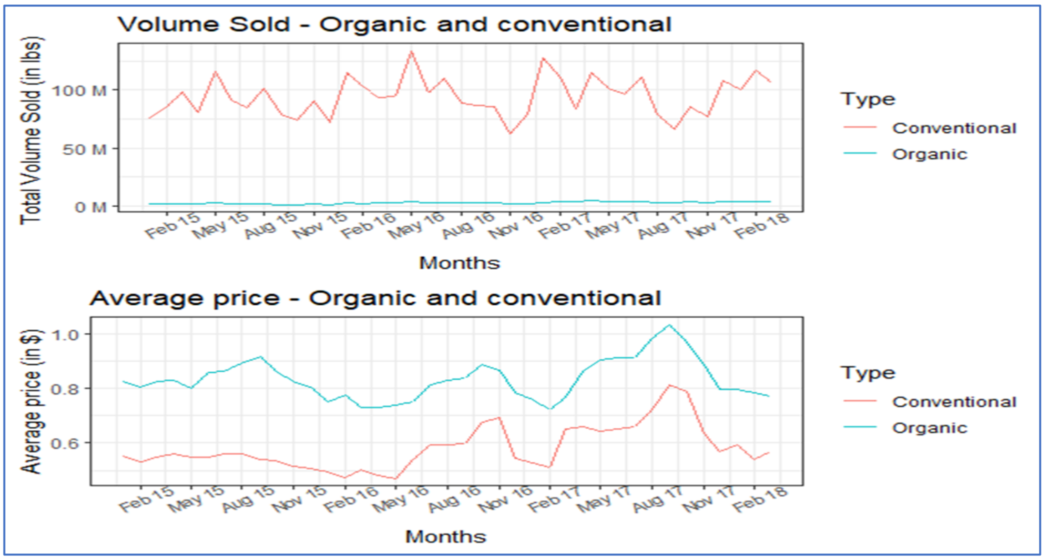
Plot 5: Time-Series trend of Avocados sold by Volume and Average Price

The above time series trend of Avocados sold by volume and average price gives us information that the highest price at which avocados are sold has been increasing year on year from 2015 to 2017 and sales fluctuate month by month. An increase in sales is observed when the price drops and vice versa.



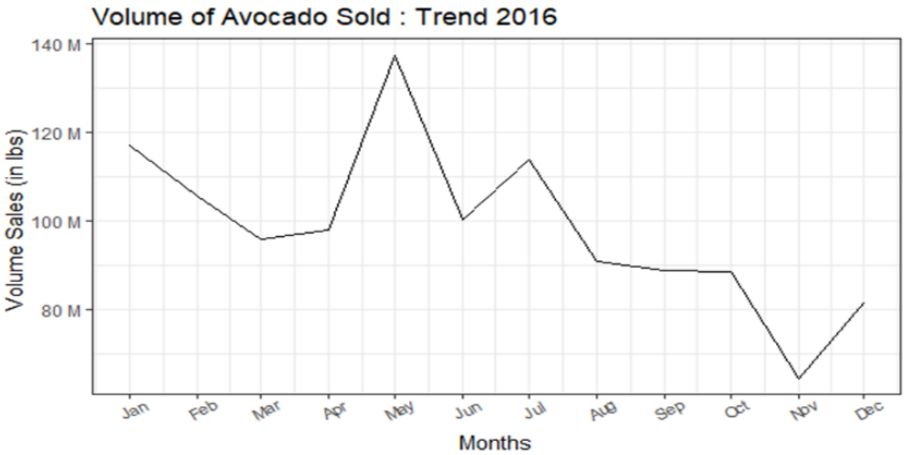
Plot 6: Time-Series trend of Avocados based on Category wise Trend and Bag Size

The above time series trend of sub-category of Avocados helps us to infer a difference in sales for different categories and bag sizes of avocados. We get to know from the visualization that Category 3 avocados are sold much less than Category 1 and Category 2. Sale of small size bags is more than large and extra-large size bags and sales for Cat3 as well as XL bags has not changed much over the years.



Plot 7: Trend of Organic and Conventional based on Volume Sold and Average Price

In above visualization we observe the trend of organic and conventional based on Volume sold and average price. The total volume and average price of conventional avocados is more as compared to organic avocados

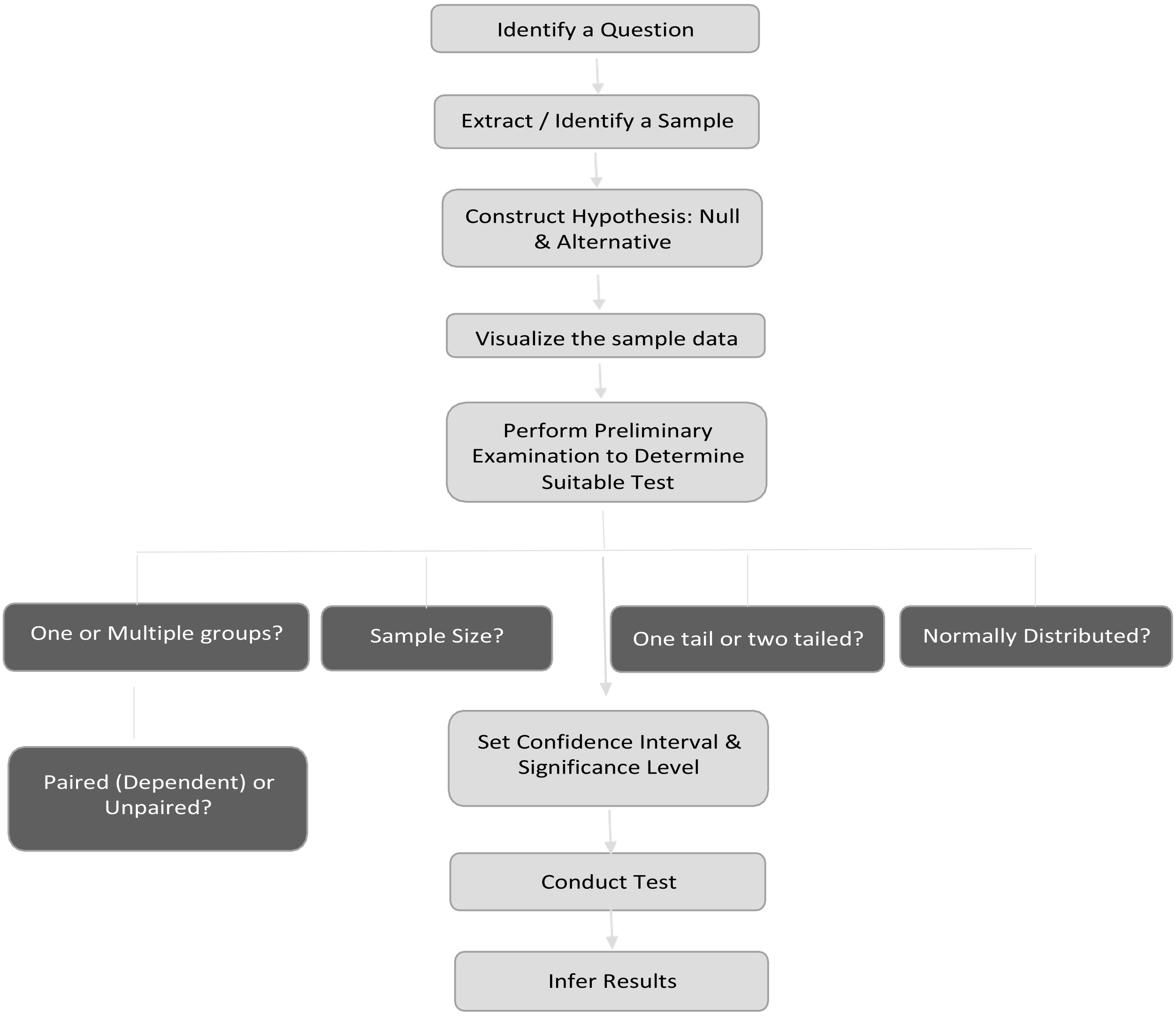


Plot 8: Time-Series trend of Volume of Avocado Sold in Year 2016

The above time-series trend of Volume of Avocado Sold explains that avocado sales vary by month based on seasonality of avocadoes as April, May and Jun is harvest season for avocados. From the graph we can visualize that the avocado sales seem to drop from August to December whereas the sales are higher in April and May.

Hypothesis Testing

Using the quantitative variables and some questions we will define hypothesis and test them statistically. The test will help us confirm if the observations from the population stated in the table below are same for any sample extracted from the data. The following flowchart illustrates the steps performed in these tests.

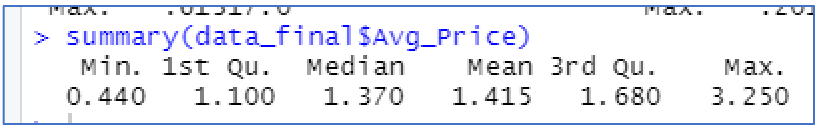


Identify a Question
Extract / Identify a Sample
Construct Hypothesis: Null & Alternative
Visualize the sample data
Perform Preliminary Examination to Determine Suitable Test
One or Multiple groups?
Sample Size?
One tail or two tailed?
Normally Distributed?
Paired (Dependent) or Unpaired?
Set Confidence Interval & Significance Level
Conduct Test
Infer Results

Test 1: One sample t-test for Average Price (two tailed)

Question 1: Is the mean average price of avocado equal to $1.4?

From the descriptive data analysis of the population dataset of 15207 observations, it was observed that the mean price of avocados sold was $1.415. We will randomly extract a sample from the total population, for testing if the any sample from this dataset complies with this finding.



```
> summary(data_final$Avg_Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.440   1.100   1.370   1.415   1.680   3.250 
```

1.1 Hypothesis

Null Hypothesis: The mean average price of avocados is equal to $1.4

Ho = ma - mb = 0

Alternative Hypothesis: The mean average price of avocados is not equal to $1.4

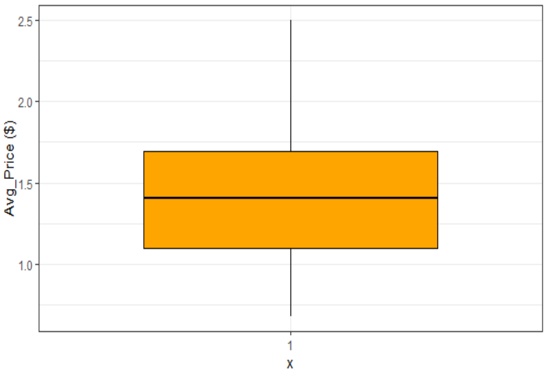
Ha = ma - mb < > 0

1.2 Sampling

For conducting the test, a sample of 50 observations is taken from the population.

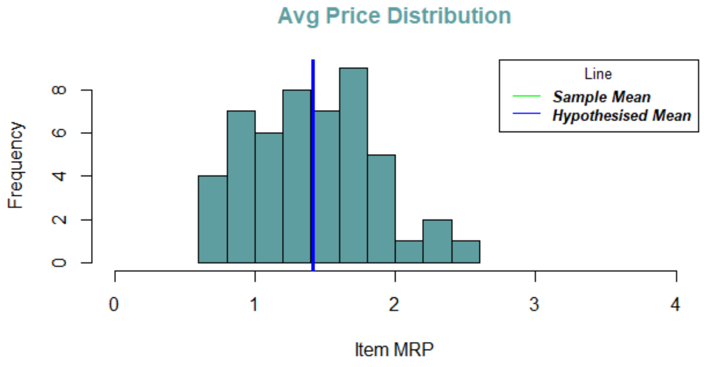
1.3 Visualizing the sample data

From the boxplot below, it represents the average price which does not have any outliers and quartiles are symmetric from the median.



Plot 9: Box Plot of sample average price of sample data

The histogram below gives a representation of distribution of avg price of the sample. The green line represents mean of sample overlaps the hypothesized mean represented by the blue line. We will further conduct tests to confirm if the sample mean varies with respect to the hypothesized mean.



Plot 10: Histogram for Avg Price distribution

1.4 Preliminary Examination to determine Test

We conduct preliminary tests to check if the data meets conditions for test assumptions and determine which test is appropriate.

- Is the sample n >30?

  Yes, n= 50

- Are we studying one group or more?

  One group, so we will use one sample t test

- Is it normally distributed?

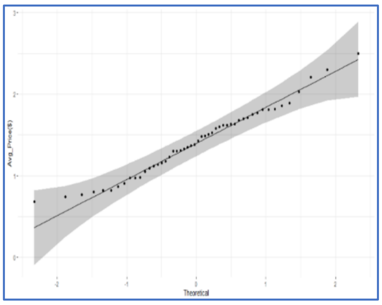
  There is an indication of data is normally distributed. We conduct Shapiro Wilk test to check normality of data.

- One tailed or two tailed?

  Since the critical area of distribution is two sided, we use a two tailed test.

1.5 Visual inspection of normality

For a visual inspection of the data normality, we use **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution. From the plot we can say that the data seems to be normally distributed.



Plot 11: Q-Q plot to check normal distribution

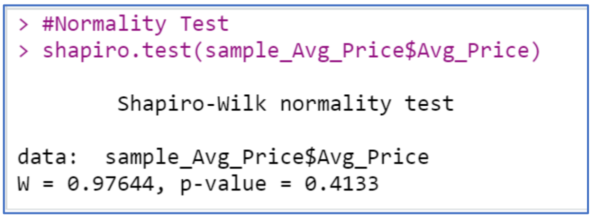
1.6 Test for Normality

We will conduct **Shapiro-Wilk test** to check if the sample data is normally distributed for average price in sample data.

Shapiro-Wilk Test for Average Price

Ho = Null hypothesis: The data is normally distributed

Ha = Alternative hypothesis: The data is not normally distributed



```
> #Normality Test
> shapiro.test(sample_Avg_Price$Avg_Price)

	Shapiro-Wilk normality test

data:  sample_Avg_Price$Avg_Price
W = 0.97644, p-value = 0.4133
```

In Shapiro-Wilk Test for average price, p-value (0.4133) > 0.05. Hence, we can say that the average price is significantly normally distributed.

1.7 One sample t-test for average price (two tailed) - Justification

Since the sample data set is assumed to be distributed normally, we conduct one sample t-test (two tailed) on it. One sample t-test is used to determine whether a sample mean is different from a specific value.

1.8 Computing

We use Confidence Interval = 95% and a = 0.05 for our test.

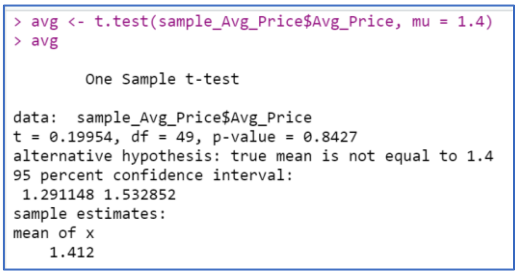
Decision rule: Reject H0 at p< 0.05

Result

The test results show that

- **t** or the **test statistic** value is 0.19954
- Degrees of freedom (df) is equal to 49

- **p-value** or the significance level 0.8427.

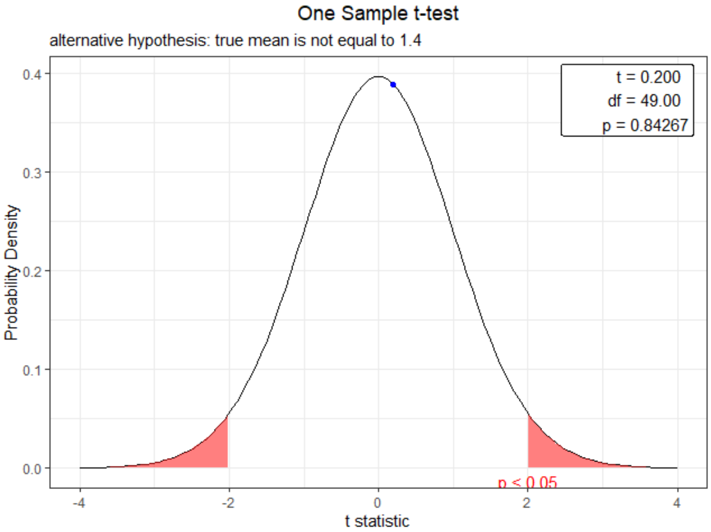


```
> avg <- t.test(sample_Avg_Price$Avg_Price, mu = 1.4)
> avg

	One Sample t-test

data:  sample_Avg_Price$Avg_Price
t = 0.19954, df = 49, p-value = 0.8427
alternative hypothesis: true mean is not equal to 1.4
95 percent confidence interval:
 1.291148 1.532852
sample estimates:
mean of x
    1.412
```

1.9 Inference

The p-value of the test is 0.8427 is more than the significance level alpha = 0.05, so we accept the null hypothesis Ho = ma - mb = 0 and conclude that the sample the mean price of avocados is equal to $1.4.

We can also observe the result on the plot below. The shaded region represents area where we reject the null hypothesis if the t value lies on it. This region is defined using the significance level alpha equal to 0.05. As we see that the blue dot does not lie on the shaded region, we do not reject the null hypothesis. We can conclude that there is enough evidence that there are 95% chances of avocados having a mean price of $1.4.

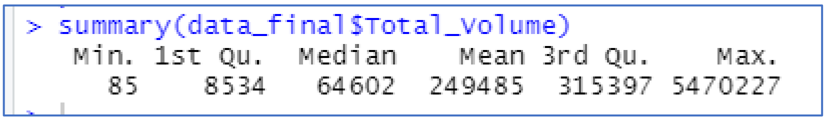


Plot 12: Paired t-test output two tailed at α =0.05

Test 2: One sample Wilcoxon signed rank test for Total Volume Sales (one tailed)

Question 2: Is the mean total volume of avocados sold greater than 200000 lbs?

From the descriptive data analysis of the population dataset of 15207 observations, it was observed that the mean volume of avocados sold was 249485 lbs. As a marketeer we would be interested in finding out if the mean of volume of avocados sold is greater than 200000 lbs all the times or only randomly. To test this, we will randomly extract a sample from the total population, for testing if the any sample from this dataset complies with this finding.



```
> summary(data_final$Total_Volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     85    8534   64602  249485  315397 5470227 
```

2.1 Hypothesis

Null Hypothesis: The mean total volume of avocado’s sold is greater than 200000 lbs.

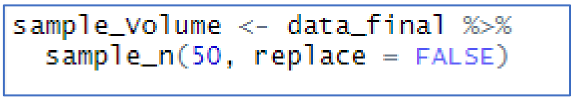
H0:m > 200000 lbs

Alternative Hypothesis: The mean total volume of avocado’s sold is not greater than to 200000 lbs.

Ha:m ≯200000 lbs

2.2 Sampling

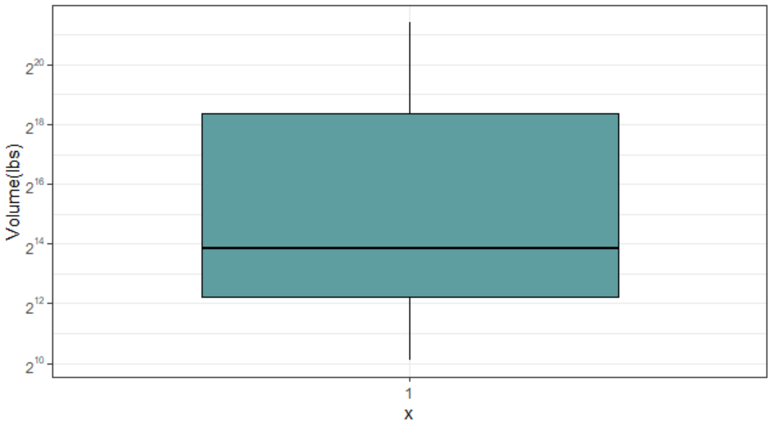
For conducting the test, a sample of 50 observations is taken from the population.



```
sample_volume <- data_final %>%
  sample_n(50, replace = FALSE)
```

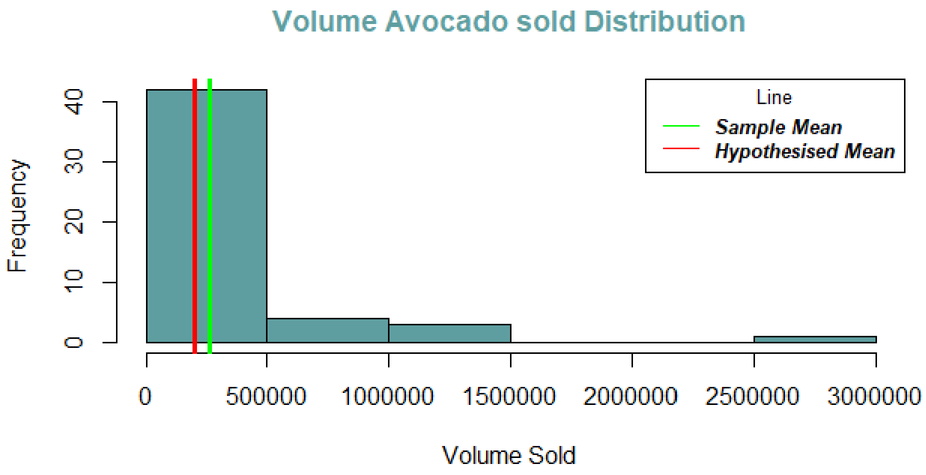
2.3 Visualizing the sample data

From the boxplot below, it represents the total volume which does not have any outliers and quartile Q1 is greater from the median.



Plot 13: Box Plot of sample total volume sales of sample data

The histogram below gives a representation of distribution of volume of avocados sold from the sample. The green line represents mean of sample which is slightly higher than the hypothesized mean represented by the red line. We will further conduct tests to confirm if the mean of avocados sold is always greater than hypothetical value 200000.



Plot 14: Histogram for Volume Sales Distribution

2.4 Preliminary Examination to determine Test

We conduct preliminary tests to check if the data meets conditions for test assumptions and determine which test is appropriate.

- Is the sample n >30?

  Yes, n= 50
- Are we studying one group or more?

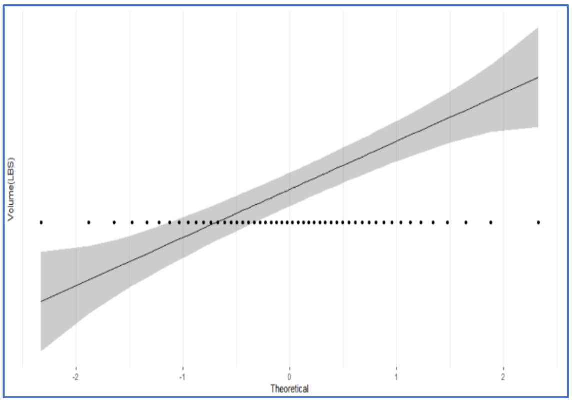
  One group, so we will use one sample t test
- Is it normally distributed?

  We conduct Shapiro Wilk test to check normality of data.
- One tailed or two tailed?

Since the critical area of distribution is one sided, we use a two tailed test.

2.5 Visual inspection of normality

For a visual inspection of the data normality, we use **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution. From the plot we can say that the data seems not normally distributed.



Plot 15: Q-Q plot to check normal distribution

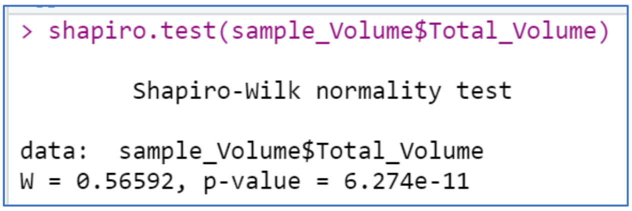
2.6 Test for Normality

We will conduct **Shapiro-Wilk test** to check if the sample data is normally distributed for volume sold in sample data.

Shapiro-Wilk Test for Average Price

Ho = Null hypothesis: The data is normally distributed

Ha = Alternative hypothesis: The data is not normally distributed.



```
> shapiro.test(sample_Volume$Total_Volume)

        Shapiro-Wilk normality test

data:  sample_Volume$Total_Volume
W = 0.56592, p-value = 6.274e-11
```

In Shapiro-Wilk Test for average price, p-value (6.274 * 10^{-11}) < 0.05. Hence, the total volume is not normally distributed.

2.7 One sample Wilcoxon signed rank test (one tailed) - Justification

As our sample data is not normally distributed, we can move forward with one sample Wilcoxon signed rank test.

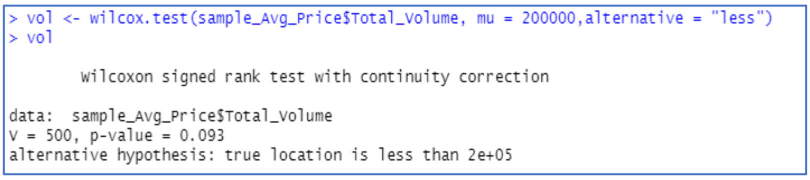
2.8 Computing

1) **With Confidence Interval = 95% and a = 0.05**

Decision rule: Reject H0 at p< 0.05

Result

The test results show that **p-value** or the significance level 0.093



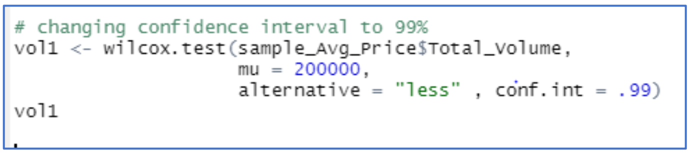
```
> vol <- wilcox.test(sample_Avg_Price$Total_Volume, mu = 200000,alternative = "less")
> vol

        Wilcoxon signed rank test with continuity correction

data:  sample_Avg_Price$Total_Volume
V = 500, p-value = 0.093
alternative hypothesis: true location is less than 2e+05
```

Inference

The p-value of the test is 0.093 is more than the significance level alpha = 0.05, so we accept the null hypothesis and conclude that the mean total volume of avocado's sold is greater than 200000 lbs. H0:m > 200000 lbs. We can conclude that there is enough evidence to say that there are 5% chances of volume avocados sold being less than 200000 lbs for a sample.

2) **With Confidence Interval = 99% and a = 0.01**

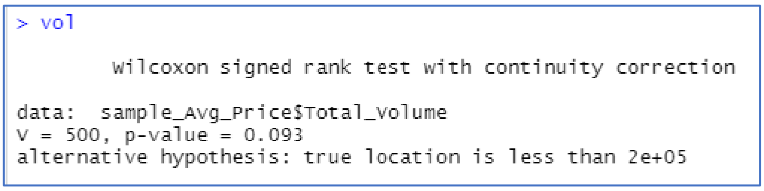
We also conducted the same test by changing the confidence interval to 99% for observing if there is any change in results. Decision rule: Reject H0 at p< 0.01



```
# changing confidence interval to 99%
vol1 <- wilcox.test(sample_Avg_Price$Total_Volume,
                    mu = 200000,
                    alternative = "less" , conf.int = .99)
vol1
```

Result

The test results show that **p-value** or the significance level 0.093



```
> vol

        Wilcoxon signed rank test with continuity correction

data:  sample_Avg_Price$Total_Volume
V = 500, p-value = 0.093
alternative hypothesis: true location is less than 2e+05
```

Inference

Even after changing the confidence interval from 95% to 99%, the p-value of the test is 0.093 is more than the significance level alpha = 0.01, so we accept the null hypothesis and conclude that the mean total volume of avocado's sold is greater than 200000 lbs. H0:m > 200000 lbs. We can conclude that there is enough evidence to say that there is only 1% chance of volume avocados sold being less than 200000 lbs for a sample.

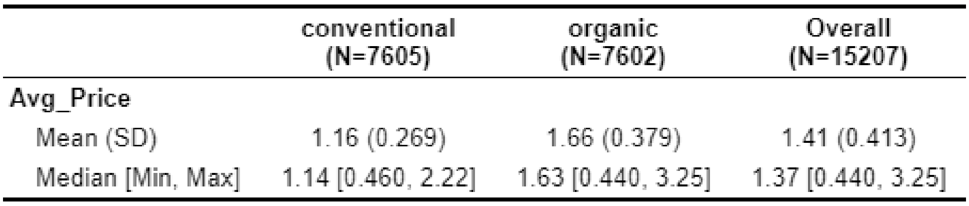
Test 3: Two sample Welch t-test (two tailed)

Question 3: Does the mean of price vary for organic and conventional avocados?

From the Table 1. below we can observe that mean price for conventional type of avocados is $1.16 while it is higher $1.66 for organic type of avocados. The minimum price for conventional avocado $1.14 is not much less than the minimum price of organic avocado $1.63. We will perform statistical test to determine this.

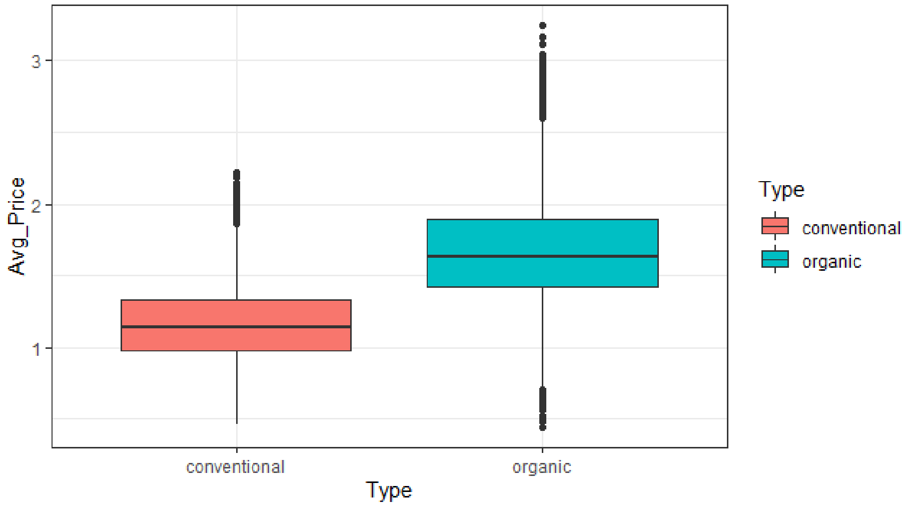
Table 2

Summary of Conventional and Organic Type Avocados



	conventional (N=7605)	organic (N=7602)	Overall (N=15207)
Avg_Price			
Mean (SD)	1.16 (0.269)	1.66 (0.379)	1.41 (0.413)
Median [Min, Max]	1.14 [0.460, 2.22]	1.63 [0.440, 3.25]	1.37 [0.440, 3.25]

We can also use the boxplot below to compare average price of conventional and organic types of avocados. It is observed that the mean price of organic type is higher than the conventional type. Some outliers are observed in both organic and conventional types of avocados sold therefore it will be prudent to confirm if the observations from the population stated in the table below are same for any sample extracted from the data.



Plot 16: Boxplot for comparing Price of Types of Avocados

3.1 Hypothesis

Null Hypothesis: There is no difference in price of organic and conventional types of avocados.

Ho = ma - mb = 0

Alternative Hypothesis: There is a difference in the price of organic and conventional types of avocados.

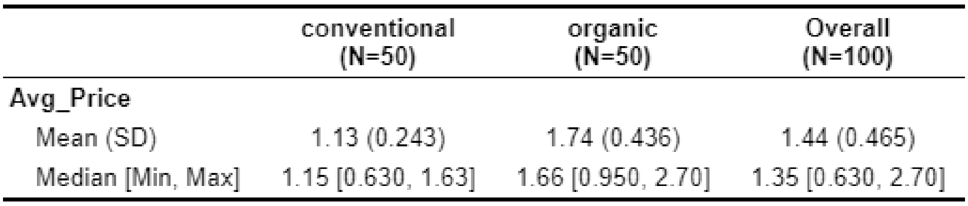
Ha = ma - mb < > 0

3.2 Sampling

For conducting the test, a sample of 100 observations, 50 each for organic and conventional type, was extracted from the total dataset and it was summarized for observation. From the sample we observe that price for conventional and organic avocado varies. Mean for organic type is $1.74 and it is $1.13 for conventional. It seems organic variety of avocados is more expensive. We will further confirm this for the sample by conducting statistical tests.

Table 3

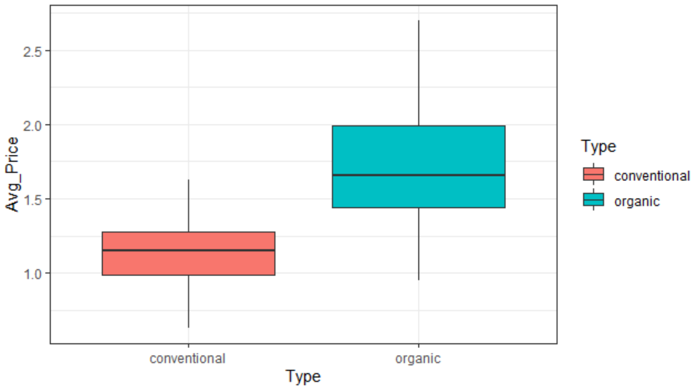
Summary of Conventional and Organic Type Avocados Sample



	conventional (N=50)	organic (N=50)	Overall (N=100)
Avg_Price			
Mean (SD)	1.13 (0.243)	1.74 (0.436)	1.44 (0.465)
Median [Min, Max]	1.15 [0.630, 1.63]	1.66 [0.950, 2.70]	1.35 [0.630, 2.70]

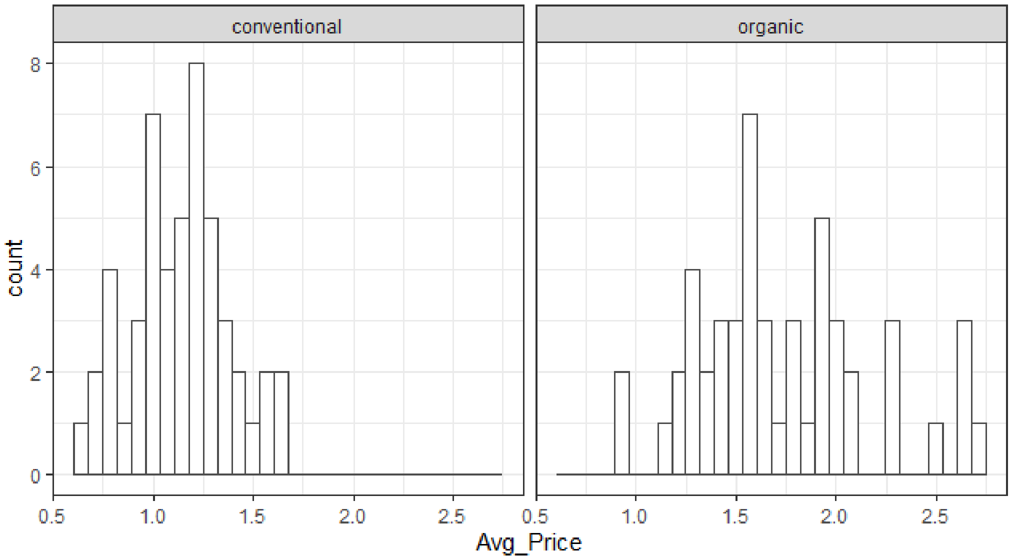
3.3 Visualizing the sample data

We visualize this data using a boxplot to check if any outliers exist. The box plot below represents no outliers exist in the data. It represents the mean comparison of price of organic and conventional types of avocados. The mean of organic seems to be visibly higher than the mean of conventional. Also. Organic variety seems to have a greater variance in price.



Plot 17: Box Plot to compare mean of price for conventional and organic

We plot a histogram to observe skewness amongst two groups. Price for most of the avocados lies between $1 to $1.5 for both varieties. We can visually infer that price variation is more for organic than for conventional. In both types the distribution seems to be somewhat closer to a normal distribution.



Plot 18: Histogram to compare price distribution for organic and conventional

3.4 Preliminary Examination to determine Test

We conduct preliminary tests to check if the data meets conditions for test assumptions and determine which test is appropriate.

- Is the sample n >30?

  Yes, n= 100 with 50 observations each for a group

- Are we studying one group or more?

  Two, organic and conventional, so we will use two sample t test

- Are two groups independent or dependent?

  Independent so we will select an unpaired test

- Is it normally distributed?

  There is an indication of data is normally distributed. We conduct Shapiro Wilk test to check normality of data.

- Is the variance of two groups equal?

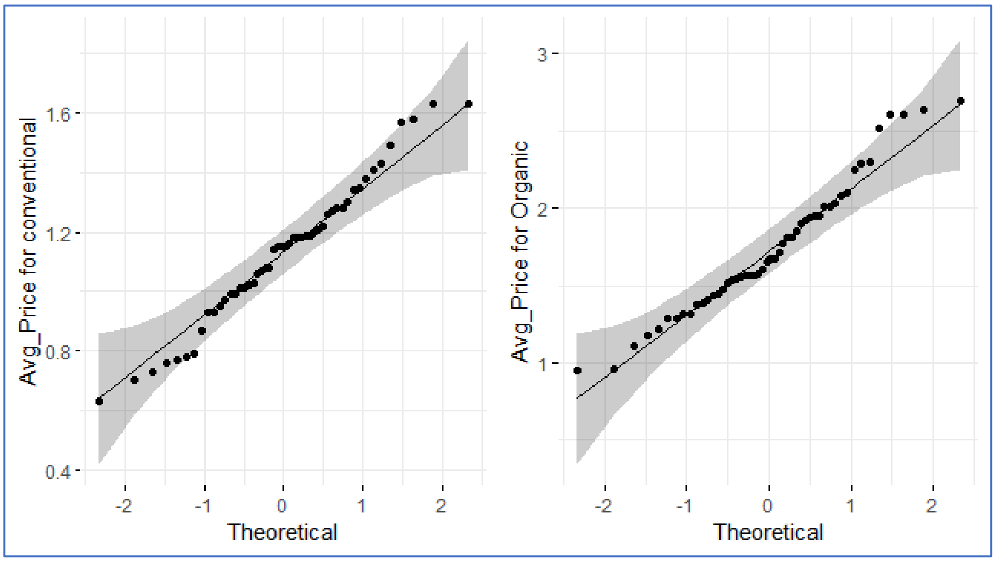
  We will check using test.

- One tailed or two tailed?

  Since the critical area of distribution is two sided, we use a two tailed test.

3.5 Visual inspection of normality

For a visual inspection of the data normality, we use **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution. From the plot we can say that the data seems to be normally distributed.



Plot 19: Q-Q plot to check normal distribution

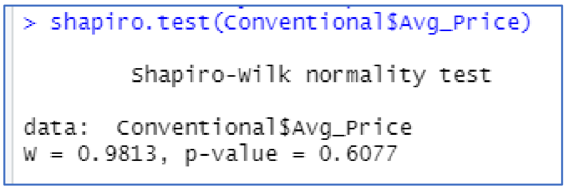
3.6 Test for Normality

We will conduct **Shapiro-Wilk test** to check if the sample data is normally distributed for both groups, organic and conventional.

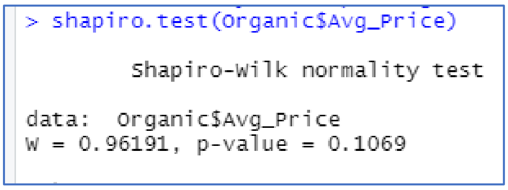
Shapiro-Wilk Test for organic and conventional avocados

Ho = Null hypothesis: The data is normally distributed

Ha = Alternative hypothesis: The data is not normally distributed



```
> shapiro.test(Conventional$Avg_Price)

        Shapiro-Wilk normality test

data:  Conventional$Avg_Price
W = 0.9813, p-value = 0.6077
```



```
> shapiro.test(Organic$Avg_Price)

        Shapiro-Wilk normality test

data:  Organic$Avg_Price
W = 0.96191, p-value = 0.1069
```

Conventional: The output of Shapiro Wilk test gives p-value =0.742 which is greater than significance level 0.05 implying that the distribution of data is not significantly different from normal distribution. We can assume that the data is normally distributed.

Organic: The output of Shapiro Wilk test gives p-value =0.2358 which is greater than significance level 0.05 implying that the distribution of data is not very different from normal distribution. We can assume that the data is normally distributed.

3.7 Test for Variance: F -test

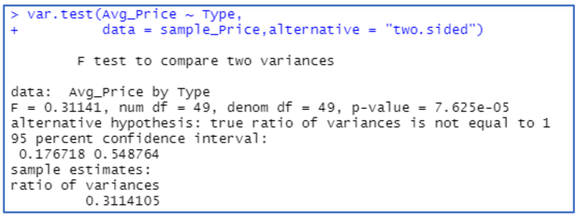
We use F-test to assess whether the variances of two populations (A and B) are equal.

Null hypothesis (H0): Variance of price for organic and conventional avocado is same.

H0: σ2A=σ2B

Alternative hypotheses (Ha): Variance of price for organic and conventional avocado is not same.

Ha: σ2A≠σ2B



```
> var.test(Avg_Price ~ Type,
+           data = sample_Price,alternative = "two.sided")

        F test to compare two variances

data:  Avg_Price by Type
F = 0.31141, num df = 49, denom df = 49, p-value = 7.625e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.176718 0.548764
sample estimates:
ratio of variances 
         0.3114105 
```

It is observed that the p-value of F-test where p value is less than the significance level 0.05. In conclusion, there is a significant difference between the two variances

3.8 Two sample Welch t-test for price based on type (two tailed) - Justification

Since the sample data is distributed normally, but variance for both groups is different, we conduct a parametric two sample t-test on it. The unpaired two-samples Welch t-test is used to compare the mean of two independent groups when a difference in variance exists.

As our data is parametric (normally distributed) and the variances of the two groups being compared is different (heteroscedasticity), we will conduct the two-samples Welch test. It is an alternative to the unpaired two-samples t-test, which can be used to compare two independent groups of samples when variance in groups is same.

3.9 Computing

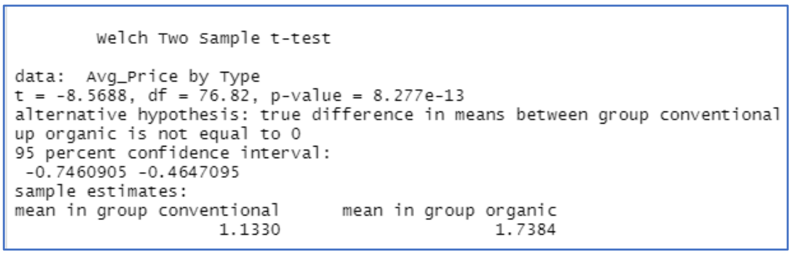
We use Confidence Interval = 95% and a = 0.05 for our test.

Decision rule: Reject H0 at p< 0.05

Result

The test results show that

- **t** or the **test statistic** value is -8.5688
- Degrees of freedom (df) is equal to 76.82
- **p-value** or the significance level 8.277e-13.



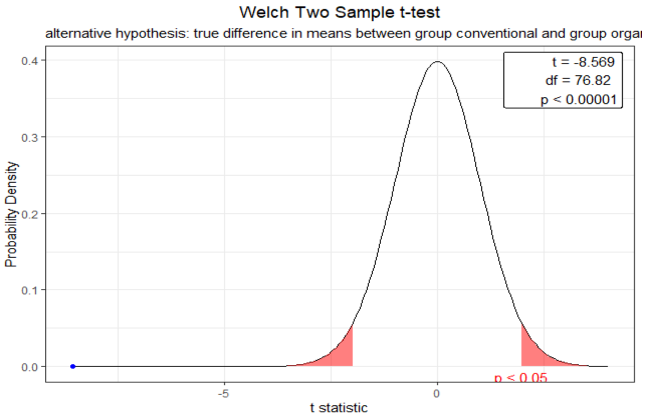
```
        Welch Two Sample t-test

data:  Avg_Price by Type
t = -8.5688, df = 76.82, p-value = 8.277e-13
alternative hypothesis: true difference in means between group conventional
up organic is not equal to 0
95 percent confidence interval:
 -0.7460905 -0.4647095
sample estimates:
mean in group conventional      mean in group organic
                    1.1330                     1.7384
```

3.10 Inference

The p-value of the test is 8.277e-13 is less than the significance level alpha = 0.05, so we reject the null hypothesis and conclude that price for organic type avocado is significantly different from conventional type or Ha = ma - mb < > 0

From the result of test, we can conclude that the sample we used was not normally distributed. As the null hypothesis is rejected, it means that whenever a sample is randomly extracted from the 15207 observations, the MRP of samples for Conventional and Organic avocados will be significantly different from each other.

We can also observe the result on the plot below. The shaded region represents area where we reject the null hypothesis if the t value lies on it. This is a two tailed test therefore the shaded region lies on both ends. This region is defined using the significance level alpha =0.05. As we see that the blue dot, or t value equal to -8.569 lies outside the curve on the shaded region, we reject the null hypothesis.



Plot 20: Welch Two Sample t-test at α =0.05

Test 4: Two sample Welch t-test (one tailed)

Question 4: Is the mean of price for conventional avocados less than organic?

Now that from the Welch t-test one tailed, we have concluded that the mean price of samples for Conventional and Organic avocados will be significantly different from each other, it would be prudent to find if the price of conventional avocados is always less than organic as observed in the preliminary examination. We will conduct the one tailed test for the same sample to find this.

4.1 Hypothesis

Null Hypothesis: The mean price of conventional avocado's is less than mean price of organic avocados.

H0:ma < mb

Alternative Hypothesis: The mean price of conventional avocado's is not less than mean price of organic avocados.

H0:ma !< mb

4.2 Two sample Welch t-test for price based on type (one tailed) Computing

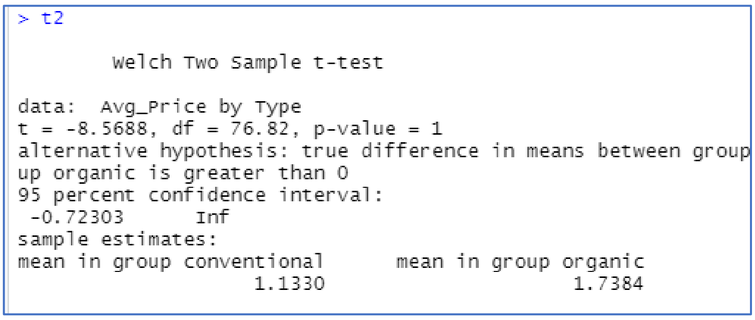
Confidence Interval is 95% and significance level = 0.05 for this test

Decision rule: Reject H0 at p< 0.05

Result

The test results show that

- **t** or the **test statistic** value is -8.5688
- Degrees of freedom (df) is equal to 76.82
- **p-value** or the significance level 1.



```
> t2

        Welch Two Sample t-test

data:  Avg_Price by Type
t = -8.5688, df = 76.82, p-value = 1
alternative hypothesis: true difference in means between group
up organic is greater than 0
95 percent confidence interval:
 -0.72303       Inf
sample estimates:
mean in group conventional      mean in group organic
                    1.1330                     1.7384
```

4.4 Inference

The p-value of the test is 1 is greater than the significance level alpha = 0.05, so we do not reject the null hypothesis and conclude that price for conventional type avocado is less than the mean price of organic type avocado or H0:ma< mb. As the null hypothesis is not rejected, it means that whenever a sample is randomly extracted from the 15207 observations, there are 95% chances that the mean price of samples for conventional avocados will be less than the mean price for organic avocados.

Findings From Hypothesis Testing

From the analysis, we can conclude that the mean average price of avocados is equal to $1.4 and mean total volume of avocado's sold is greater than 200000 lbs. While the price and volume sold information can be helpful in building revenue models, the volume sold information can also be used in planning inventory. This observation leads to a follow up question for the companies; Will the volume of avocado's sold falls below 200000 lbs if the price of avocados is slightly increased above $1.4?

We have an important finding that avocados sold, under the category of conventional and organic, have a significantly different price. The organic avocados are costlier than the conventional avocados. As we observed in the exploratory analysis that conventional avocados are sold much more than conventional avocados, companies can focus on marketing organic avocados. This can evidently prove helpful in improving revenues for companies.

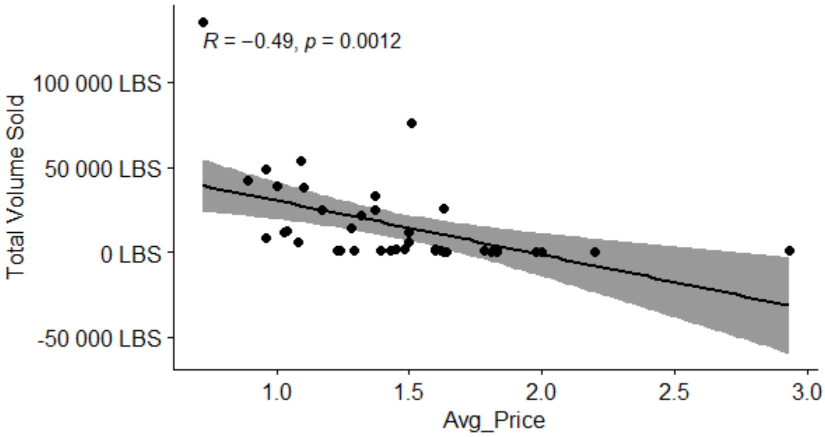
Correlation

Correlation Two Variables

Question 1: Is there a relationship amongst total volume sales and average price?

Visualizing the data

From the plot we can observe a negative linear correlation of total volume sales and average price. It is observed that as the average price of items is increasing, lesser avocados are being sold. R coefficient is equal to -0.49.



Plot 21: Scatter plot for average price and Total Volume Sales

Correlation

Correlation in R can be produced using cor()function. The following values are obtained.

Table 2: Correlation Table

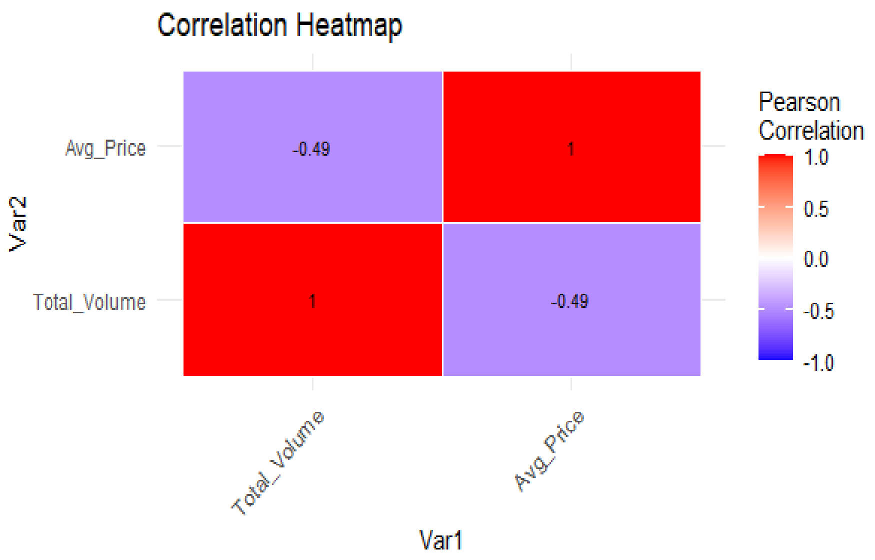
	Total_Volume Sales	Avg_Price
Total_Volume Sales	1	
Avg_Price	-0.49	1

Correlation Heatmap

To understand the relationship better, we plot a correlation matrix. **Positive correlations** are displayed in red and **negative correlations** in blue color. Color intensity is proportional to the **correlation coefficients**. In the right side the legend color shows the correlation coefficients and the corresponding colors.

Observations

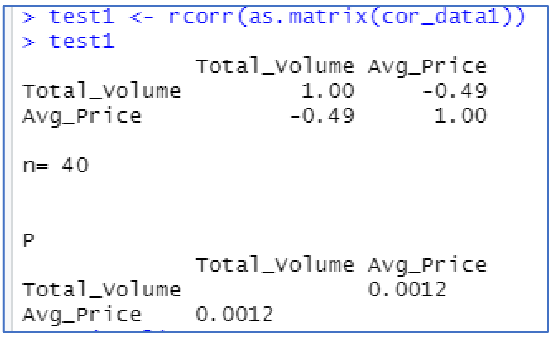
A negative correlation with very small correlation coefficient is observed between Average price and Total volume sales.



Plot 22: Correlation Heatmap for Total Volume Sales and Average Price

Correlation matrix with significance levels (p-value)

The function rcorr() can be used to compute the significance levels for Pearson . The output gives the correlation matrix n and corresponding p values indicating the significance levels of correlations.



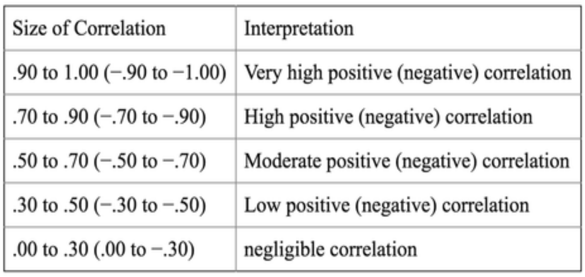
```
> test1 <- rcorr(as.matrix(cor_data1))
> test1
             Total_Volume Avg_Price
Total_Volume         1.00     -0.49
Avg_Price           -0.49      1.00

n= 40


P
             Total_Volume Avg_Price
Total_Volume              0.0012
Avg_Price    0.0012
```

Row	Column	cor	p
Total_Volume_Sales	Avg_Price	-0.49	0.0012

Observation:

- The **p-value** of the test is 0.0012, which is less than the significance level alpha = 0.05. We can conclude that there is enough evidence to suggest that average price and total volume sales are significantly **correlated** with a correlation coefficient of -0.49.

The correlation observed is a **low negative correlation** based on the value = -0.49 of correlation coefficient. We can say that more avocados are sold when price is less.



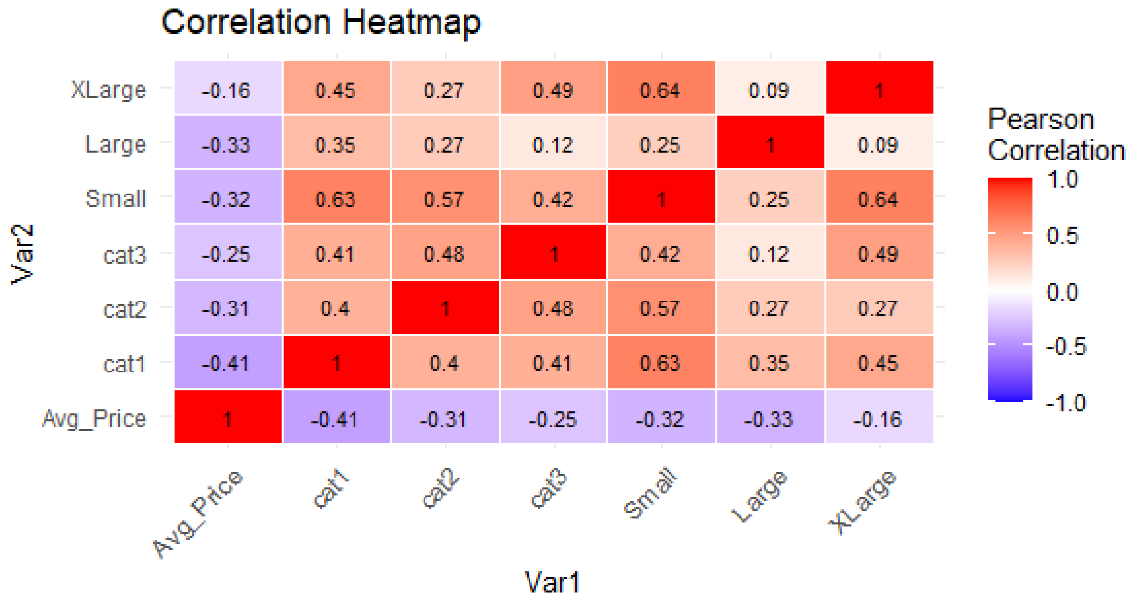
Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Correlation with multiple variables

Question 2: Is there a relationship in total volume sales from different bag sizes, different categories and average price?

Correlation Heatmap

To understand the relationship better, we plot a correlation matrix. **Positive correlations** are displayed in red and **negative correlations** in blue color. Color intensity is proportional to the **correlation coefficients**. In the right side the legend color shows the correlation coefficients and the corresponding colors. The pl



Plot 23: Correlation Heatmap for Total Volume Sales in different bag sizes, categories and Average Price

Observations

- A negative correlation with small correlation coefficient is observed between Average price and Total volume sales for all bag sizes.
- The correlation is very less in XL size bags (-0.16) while it is more significant in small (-0.32) and large size bags (0.33).
- The avg price and sales correlation is more significant in Cat (0.41) as compared to Cat2(0.31) and Cat3(0.25).
- Also, with a correlation coefficient of 0.63 we can say that more small sized Cat1 avocados are being sold.

Regression

Simple Linear Regression

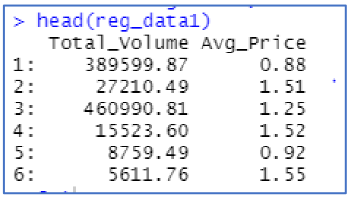
Question 2: How does Average price influence the total volume sales?

A significant but **low negative correlation** was observed between price and volume sales based on the value = -0.49 of correlation coefficient. Using regression analysis, we can now find how change in average price can influence change total volume sales. For the analysis,

- **Dependent Variable**: Total Volume Sales
- **Independent Variable**: Average Price

Sampling

We will randomly extract a sample from the total population for testing if change in MRP influence change in number of items sold. For conducting the test, a sample of 50 observations was extracted from the total dataset. We can then observe the summary of item MRP and number of items sold in the sample dataset to get an idea of min, max, mean and median.



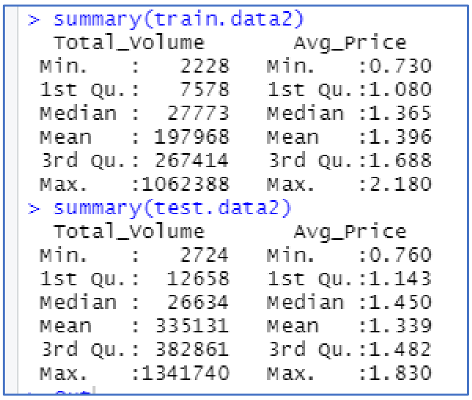
```
> head(reg_data1)
   Total_Volume Avg_Price
1:    389599.87      0.88
2:     27210.49      1.51
3:    460990.81      1.25
4:     15523.60      1.52
5:      8759.49      0.92
6:      5611.76      1.55
```

Partition data

We then partition the data into **test and train** groups. The train data has 32 values while test has 8. We will build the regression model on Tarin data set while use the Test dataset to predict values. From the test data set we will use the predictor and outcome variables for our analysis.

- **Predictor Variable:** Average price in test data set

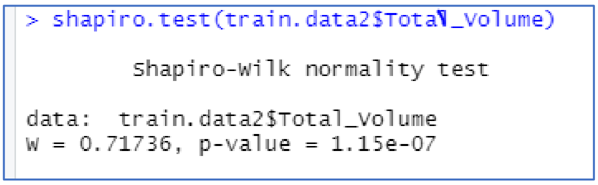
- **Outcome:** Total volume sales



```
> summary(train.data2)
  Total_Volume        Avg_Price    
 Min.   :   2228   Min.   :0.730  
 1st Qu.:   7578   1st Qu.:1.080  
 Median :  27773   Median :1.365  
 Mean   : 197968   Mean   :1.396  
 3rd Qu.: 267414   3rd Qu.:1.688  
 Max.   :1062388   Max.   :2.180  
> summary(test.data2)
  Total_Volume        Avg_Price    
 Min.   :   2724   Min.   :0.760  
 1st Qu.:  12658   1st Qu.:1.143  
 Median :  26634   Median :1.450  
 Mean   : 335131   Mean   :1.339  
 3rd Qu.: 382861   3rd Qu.:1.482  
 Max.   :1341740   Max.   :1.830  
```

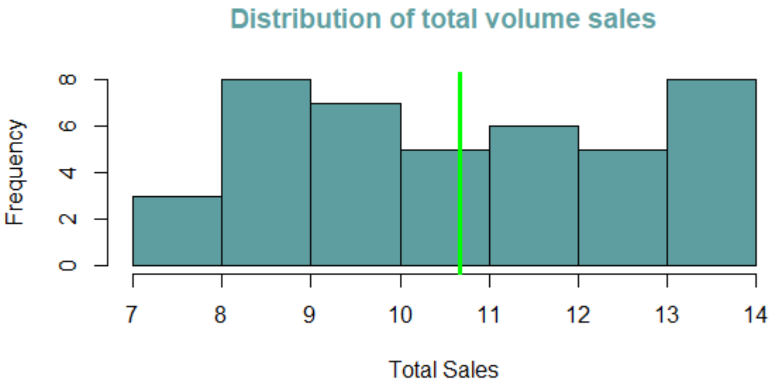
Check for Assumptions

1. **Independence of observations** (no autocorrelation): Because we only have one independent variable and one dependent variable, we don't need to test for any hidden relationships among variables.
2. **Normality:** To check whether the dependent variable Items sold follow a normal distribution, we conduct a Shapiro Wilk test.



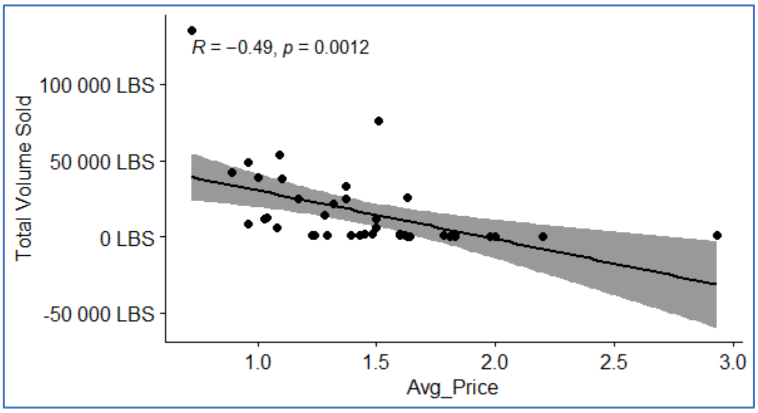
```
> shapiro.test(train.data2$Total_Volume)

        Shapiro-Wilk normality test

data:  train.data2$Total_Volume
W = 0.71736, p-value = 1.15e-07
```

The output of Shapiro Wilk test gives p-value less than significance level 0.05 implying that the distribution of data is different from normal distribution. Therefore, we shall convert it into normal distribution and observe in the histogram.



Plot 24: Histogram for distribution

Linearity: The relationship between the independent and dependent variable must be linear. We can test this visually with a scatter plot to see if the distribution of data points could be described with a straight line. We observed a **low negative correlation** with the correlation coefficient = -0.49. The relationship looks roughly linear, so we can proceed with the linear model.



Plot 25: Scatter plot for Average price and Total volume sales

Linear regression

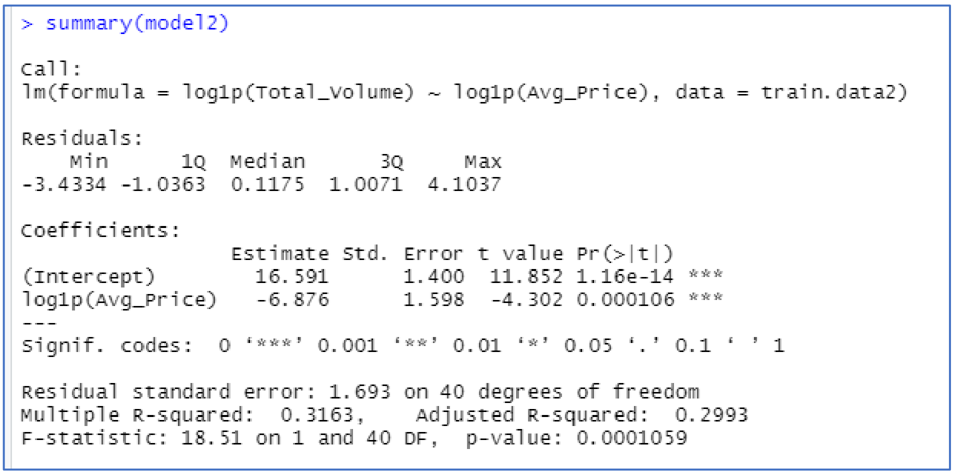
Hypothesis

H0: $\rho = 0$ there is no correlation between the average price and total volume sales.

H1: $\rho \neq 0$ there is a significant correlation between the average price and total volume sales.

Test

In R lm() function can be used to makes the linear model. Summary of the model gives us the following output.



```
> summary(model2)

Call:
lm(formula = log1p(Total_Volume) ~ log1p(Avg_Price), data = train.data2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4334 -1.0363  0.1175  1.0071  4.1037 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        16.591      1.400  11.852 1.16e-14 ***
log1p(Avg_Price)   -6.876      1.598  -4.302 0.000106 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.693 on 40 degrees of freedom
Multiple R-squared:  0.3163,	Adjusted R-squared:  0.2993 
F-statistic: 18.51 on 1 and 40 DF,  p-value: 0.0001059
```

Observations:

1. The estimates (**Estimate**) for the model parameters – the value of the y-intercept is 16.591 and the estimated effect of item MRP is -6.876
2. The residual standard error is 1.693.
3. The test statistic (**F value**, in this case the *t*-statistic) is 18.51
4. The *p*-value is 0.0001059

Model Equation: Total Volume Sale = 16.591- 6.876 *Avg Price

Inference:

- the F-statistic value = 18.51 evaluates that there is significant association between average price and total volume sales.
- From these results, as p = 0.0001059 which is less than 0.05 we can say that there is strong evidence **against** null hypothesis indicating a **significant correlation between** average price and Total volume sales. This means if the item average price is increased the sales will decrease.

Model Accuracy

Now that we have identified that average price is significantly associated to the sale, we will continue the diagnostic by checking how well the model fits the data. This process is also referred to as the goodness-of-fit

R-squared and Adjusted R-squared

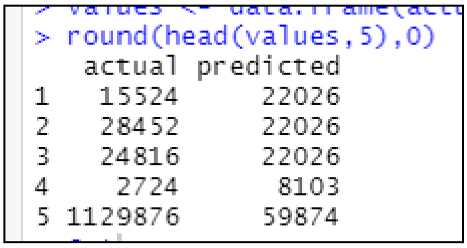
The R-squared (R2) ranges from 0 to 1 and represents the proportion of variation in the outcome variable that can be explained by the model predictor variables. The R2 measures, how well the model fits the data. The higher the R2, the better the model. Adjusted R-squared, which is a penalized R2 for a higher number of predictors. A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our example, adjusted R square is equal to 0.2993 which is low and indicates that only 29.93% of the total variation is explained by the regression line using the independent variable.

Predictions from simple linear Regression

Based on the model, we will use new predictor values of average price and total sales.

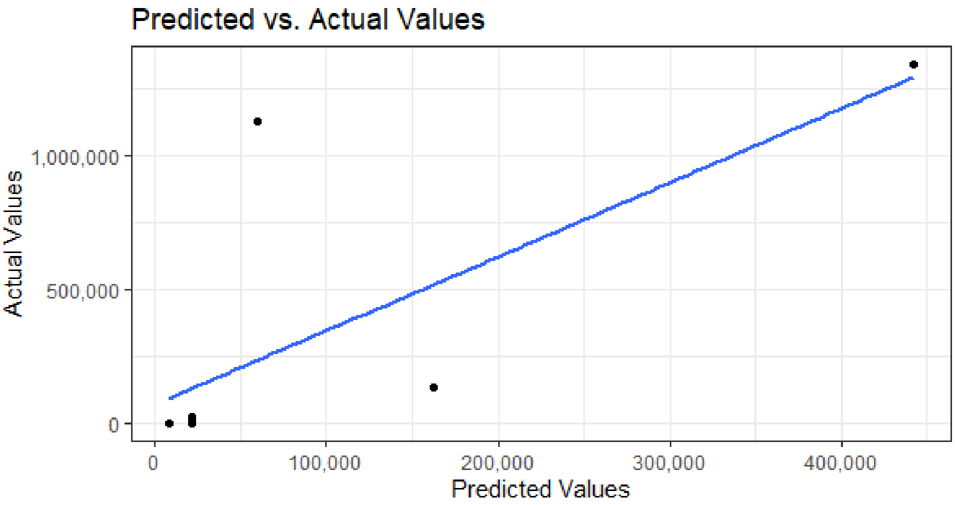
We run the model on values in test data that we had earlier partitioned from the sample. We then observe the actual values vs the predicted values. A difference in actual and predicted values will exist due to low R2 = 29%.



```
> values <- data.frame(actu
> round(head(values,5),0)
   actual predicted
1   15524     22026
2   28452     22026
3   24816     22026
4    2724      8103
5 1129876     59874
```

Actual vs Predicted Visualization

The x-axis displays the predicted values from the model and the y-axis displays the actual values from the dataset. The diagonal line in the middle of the plot is the estimated regression line. Since each of the data points lies fairly close to the estimated regression line, this tells us that the regression model does a moderate job of fitting the data.



Plot 26: Scatter plot with regression line for Actual vs Predicted Value

Multilinear Regression

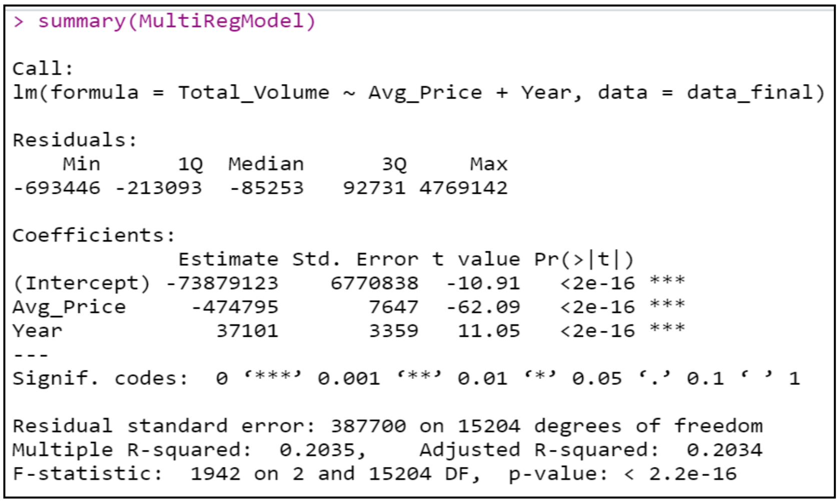
Question 1: Does the Average price factor effects Volume Sales of avocados each year?

- Dependent Variable: Total Volume
- Predictor Variable: Average Price, Year

Model Equation:

Y(Average_Price) = B0+(Total_Volume) X1+(Year) X2

Observation

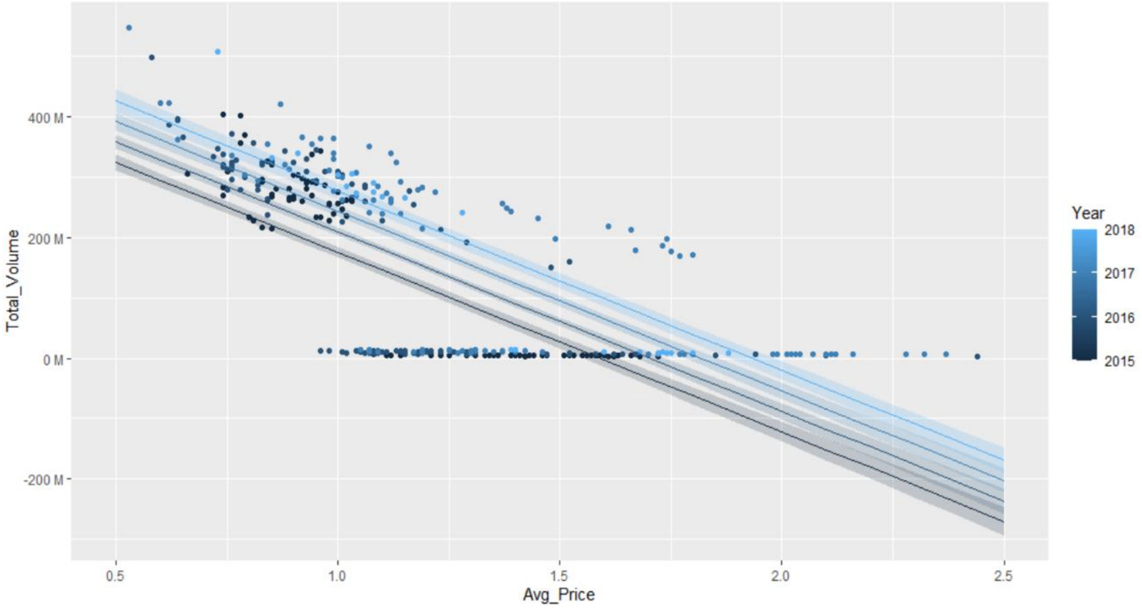


```
> summary(MultiRegModel)

Call:
lm(formula = Total_Volume ~ Avg_Price + Year, data = data_final)

Residuals:
    Min      1Q  Median      3Q     Max 
-693446 -213093  -85253   92731 4769142 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -73879123    6770838  -10.91   <2e-16 ***
Avg_Price     -474795       7647  -62.09   <2e-16 ***
Year            37101       3359   11.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 387700 on 15204 degrees of freedom
Multiple R-squared:  0.2035,	Adjusted R-squared:  0.2034 
F-statistic:  1942 on 2 and 15204 DF,  p-value: < 2.2e-16
```

The multiple regression model above we get a lower accuracy of 20.34% obtained from adjusted R-squared. We conclude that total volume sales do not has much effect by factors average price and year.

The Model Equation obtained:

Total Volume = (-73879123) – 474795*Avg_Price + 37101*Year

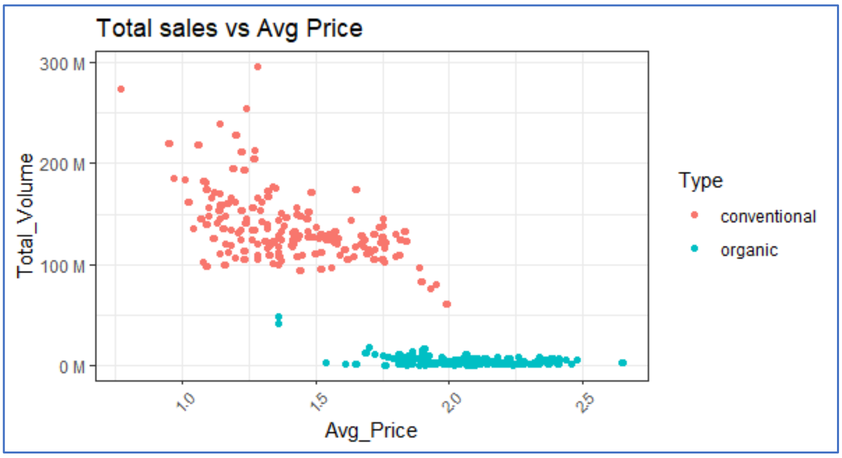


Plot 27: Scatter plot between Total Volume Sales and Average Price by each year

From the above visualization we can depict that the with increase in total volume sales there is a decrease in average price of avocados which is same in each year consecutively as we have inferred from over predicted model there is no effect in volume sales based on year.

Question 2: How does average price influence the total sales for different types?

In the EDA we used the jitter plot, to visualize the relationship of total sales and average price based on the type of avocado. Though we could observe higher sales at lower price, the difference was not evident in both types. We will use multilinear analysis to identify this. We will use New York region for this analysis.



Plot 28: Jitter Plot for average price vs total sales based on avocado type

A significant but low negative correlation was observed between average price and total sales based on the value = -0.49 of correlation coefficient. Using linear regression analysis, we identified that there is strong evidence indicating a significant correlation them. Now using multi linear regression we will analyze impact of type of avocado on it.

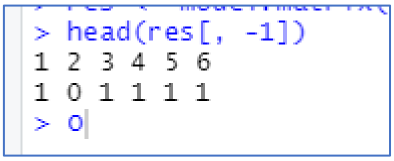
Partition Data

We then partition the data into test and train groups. We will build the regression model on Tarin data set while use the Test dataset to predict values. From the test data set we will use the predictor and outcome variables for our analysis.

- **Dependent Variable:** Total volume sales
- **Predictor Variable:**
    - Average price (Numeric and Continuous)
    - Type: (Categorical): Organic, Conventional

Dummy Variables

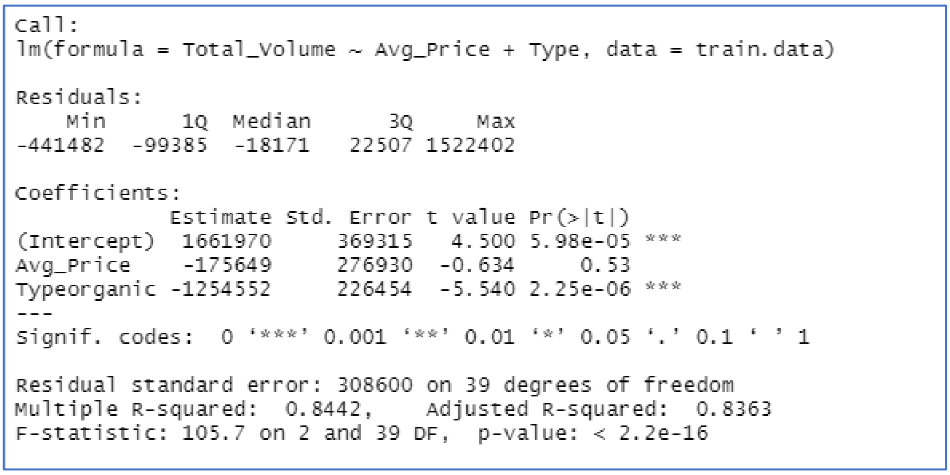
This dummy coding is automatically performed by R. For demonstration purpose, you can use the function model.matrix() to create a contrast matrix for a factor variable.



```
> head(res[, -1])
1 2 3 4 5 6
1 0 1 1 1 1
> 0
```

Compute the model

In R lm() function can be used to makes the linear model. First, we observe the output of multiple regression for Total volume sales and average price based on type.



```
Call:
lm(formula = Total_Volume ~ Avg_Price + Type, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-441482  -99385  -18171   22507 1522402 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1661970     369315   4.500 5.98e-05 ***
Avg_Price    -175649     276930  -0.634     0.53    
Typeorganic -1254552     226454  -5.540 2.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308600 on 39 degrees of freedom
Multiple R-squared:  0.8442,	Adjusted R-squared:  0.8363 
F-statistic: 105.7 on 2 and 39 DF,  p-value: < 2.2e-16
```

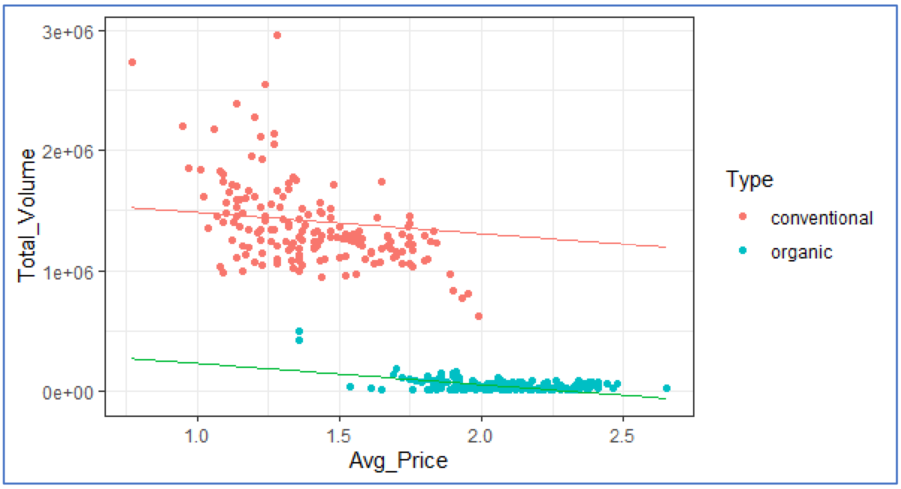
Observations:

1. The estimates (**Estimate**) for the model parameters 1661970
2. The residual standard error is 0.463.
3. Adjusted R square is 0.83
4. The F statistic value is 105.
5. The *p*-value is p-value: 2.2e-16

Model Equation:

Organic: y= -175649*x1+ 1661970

Conventional: y= -175649*x2 + 1661970 – 1254552



Plot 29: Multilinear Regression of items sold with MRP and outlet type

Inference

From the results, as p = 2.2e-16 which is less than 0.05 we can say that there is strong evidence **against** null hypothesis indicating a **significant correlation between** Items sold and all other predictor variables that is total sales, average price and type of avocados. Also, we can notice that p value for all individual predictor variables is also less than < 0.05.

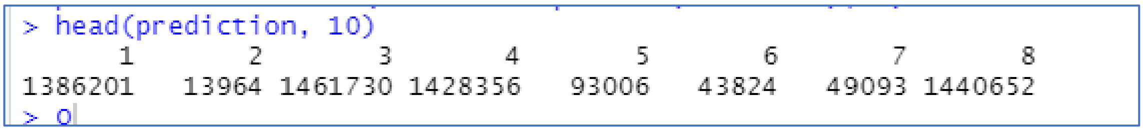
Model Accuracy

Now that we have identified that Item MRP and type of store are significantly associated to the items sold, we will continue the diagnostic by checking the goodness of fit.

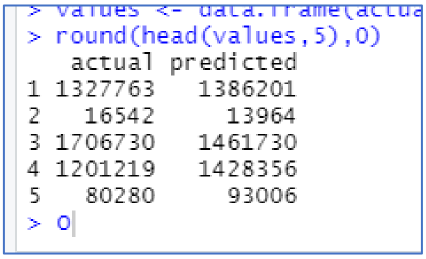
- **Residual standard error (RSE):** In our example, the RSE = 308600, meaning that the observed sales values deviate from the predicted values by approximately 308600 units in average. This corresponds to an error rate of 308600/mean(total sales) = 308600/772024= 40%.
- **R-squared and Adjusted R-squared:** In our example, adjusted R square is equal to 83.63 which indicates that only 83% of the total variation is explained by the regression line using the independent variable.

Predictions from multiple linear Regression

Based on the model, we will use new predictor values to observe Items sold. We run the model on test data that we had earlier partitioned from the sample. We then observe the actual values vs the predicted values.



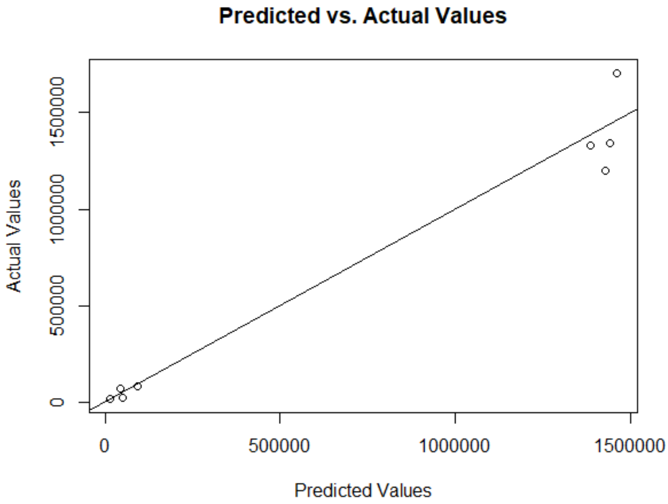
```
> head(prediction, 10)
      1       2       3       4       5       6       7       8
1386201   13964 1461730 1428356   93006   43824   49093 1440652
> 
```



```
> values <- data.frame(actua
> round(head(values,5),0)
   actual predicted
1 1327763   1386201
2   16542     13964
3 1706730   1461730
4 1201219   1428356
5   80280     93006
> 
```

Actual vs Predicted Visualization

The x-axis displays the predicted values from the model and the y-axis displays the actual values from the dataset. The diagonal line in the middle of the plot is the estimated regression line. Since each of the data points lies fairly close to the estimated regression line, this tells us that the regression model does a good job of fitting the data.



Plot 30: Predicted vs Actual Item sold value

Conclusion

From avocados analysis performed, we obtained information from the avocado dataset that the average price of avocados remains in the range of $0.4 to $3.2. Avocadoes were sold in various regions, we found that Los Angeles sells highest volume of avocados followed by New York. We observed the pattern between average price and year where we infer that the highest price at which avocados are sold has been increasing year on year from 2015 to 2017. Depending upon month wise sales, Avocado sales are higher in April and May which is also it's harvest season and drop from August to December is observed. Depending upon types of bags sold, small size bags are preferred more than large and extra-large size bags. Based on type of avocados, price as well as total volume for organic type avocado is significantly higher from conventional type which also obtained from the result of two sample t-test hypothesis conducted. From regression analysis, we found that an increase in sales is observed when the price drops and vice versa. A multiple regression we can say with 83% accuracy that total volume sales are dependent on average price depending on the type of avocado. We can therefore use average price and type of avocado values to predict future sales of avocados with 83% accuracy.

References

Regional composite reports. Hass Avocado Board. (2021, September 10). Retrieved December 15, 2021, from https://hassavocadoboard.com/regional-composite-reports/ .
Introduction to hypothesis testing in R - learn every concept from scratch! DataFlair. (2021, August 25). Retrieved December 15, 2021, from https://data-flair.training/blogs/hypothesis-testing-in-r/

Significance test for linear regression. Significance Test for Linear Regression | R Tutorial. (n.d.). Retrieved December 15, 2021, from http://www.r-tutor.com/elementary-statistics/simple-linear-regression/significance-test-linear-regression .

Grolemund, H. W. and G. (n.d.). *R for data science*. 7 Exploratory Data Analysis | R for Data Science. Retrieved December 15, 2021, from https://r4ds.had.co.nz/exploratory-data-analysis.html .

Bevans, R. (2020, October 26). *An introduction to multiple linear regression*. Scribbr.

Retrieved December 15, 2021, from https://www.scribbr.com/statistics/multiple-linear-regression/ .

One-sample T-test in R. STHDA. (n.d.). Retrieved December 15, 2021, from http://www.sthda.com/english/wiki/one-sample-t-test-in-r

Unpaired two-samples T-test in R. STHDA. (n.d.). Retrieved December 15, 2021, from http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r .

Janiobachmann. “Price of Avocados || Pattern Recognition Analysis.” Kaggle, Kaggle, 29 Apr. 2019, https://www.kaggle.com/janiobachmann/price-of-avocados-pattern-recognition-analysis.

Kabacoff, R. I.(2011). R in action: Data analysis and graphics with r. Manning Publications Co.

“T-Distribution Table (One Tail and Two-Tails).” Statistics How To, 1 June 2021, https://www.statisticshowto.com/tables/t-distribution-table/.

“Unpaired Two-Samples Wilcoxon Test in R.” STHDA, http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r.

“One-SampleT-TestinR.”STHDA, http://www.sthda.com/english/wiki/one-sample-t-test-in-r.

Bluman, A. G. (2019). Elementary statistics: A step by step approach: A brief version (Seventh). McGraw-Hill. Retrieved November 12, 2021,from https://bmalone.weebly.com/uploads/2/2/3/9/22391186/bluman_statistics_book.pdf.