Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

Study Objective:

The aim of this study is to uncover trends in avocado pricing. The analysis will focus on the following aspects:

1. Average price based on type: Conventional vs. Organic
2. Seasonal fluctuations: Identifying months with significantly higher or lower prices
3. Geographical variations: Determining regions with consistently lower or higher prices.

Description:

The Kaggle dataset provides insights into the avocado market across various cities and regions in the United States. It contains weekly retail data including volume and price of Hass avocados from January 2015 to March 2018.

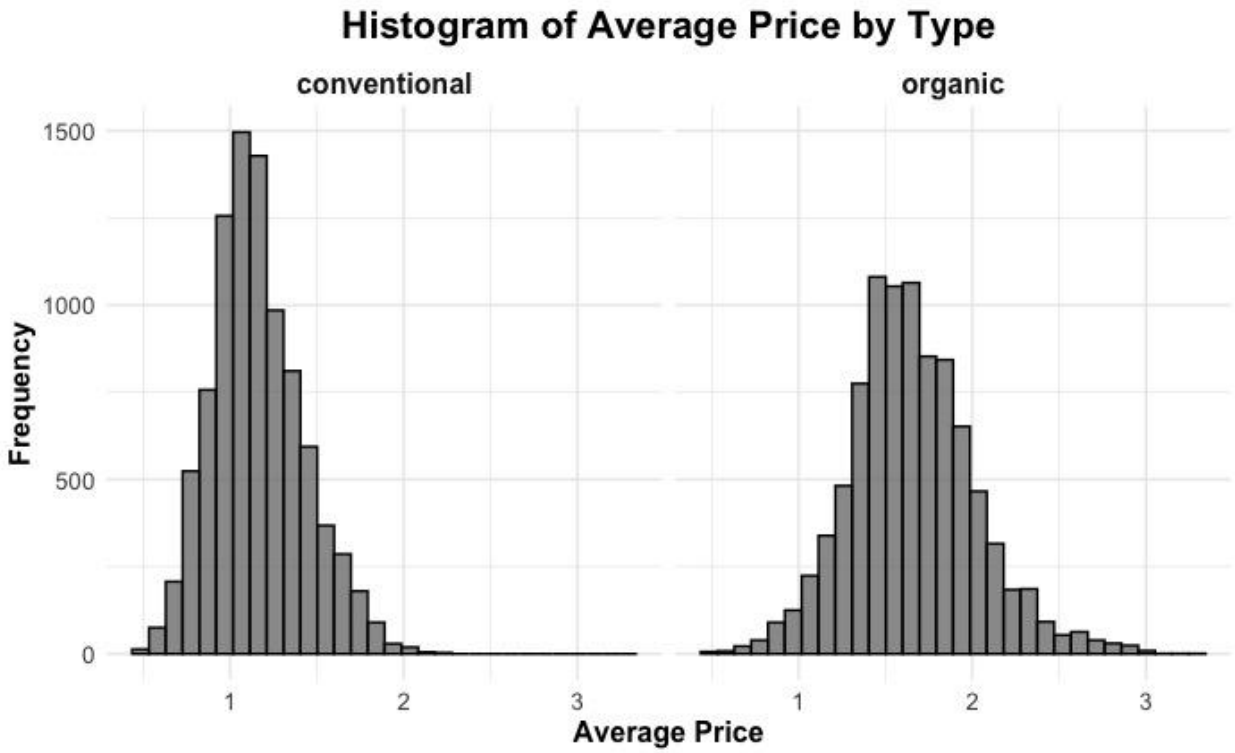
Key columns in the dataset include:

1. Date: The date of the observation.
2. AveragePrice: The average price of a single avocado.
3. Total Volume: Total number of avocados sold.
4. Type: Conventional or organic.
5. Region: The city or region of the observation.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

Exploratory Data Analysis:

Average price is approximately normally distributed within each type.

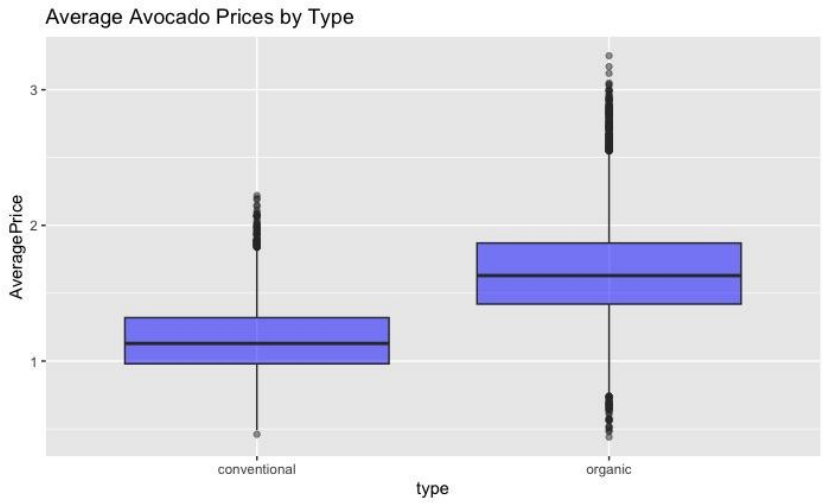


The average price of both conventional and organic avocados are normally distributed. Distribution of conventional avocados is less spread out than the distribution of organic avocados.

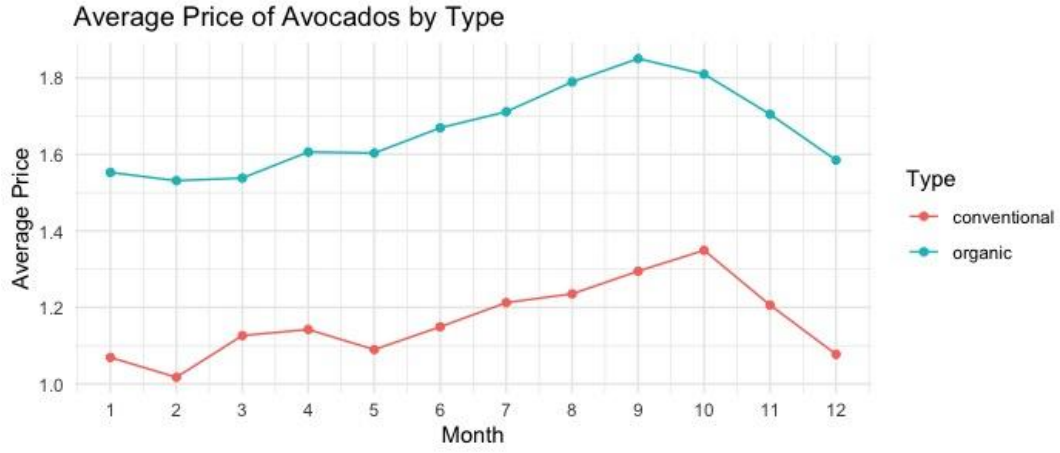
```
> # Get a summary of the average price for each type
> summary_by_type <- by(avocado_data$AveragePrice, avocado_data$type, summary)
> # Print the summary information
> print(summary_by_type)
avocado_data$type: conventional
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.460   0.980   1.130   1.158   1.320   2.220
------------------------------------------------------------------------
avocado_data$type: organic
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.440   1.420   1.630   1.654   1.870   3.250
```

There is a significant difference in the mean and median price of the two avocado types. As expected, the mean avocado price of organic avocados is much higher than that of it's counterpart.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

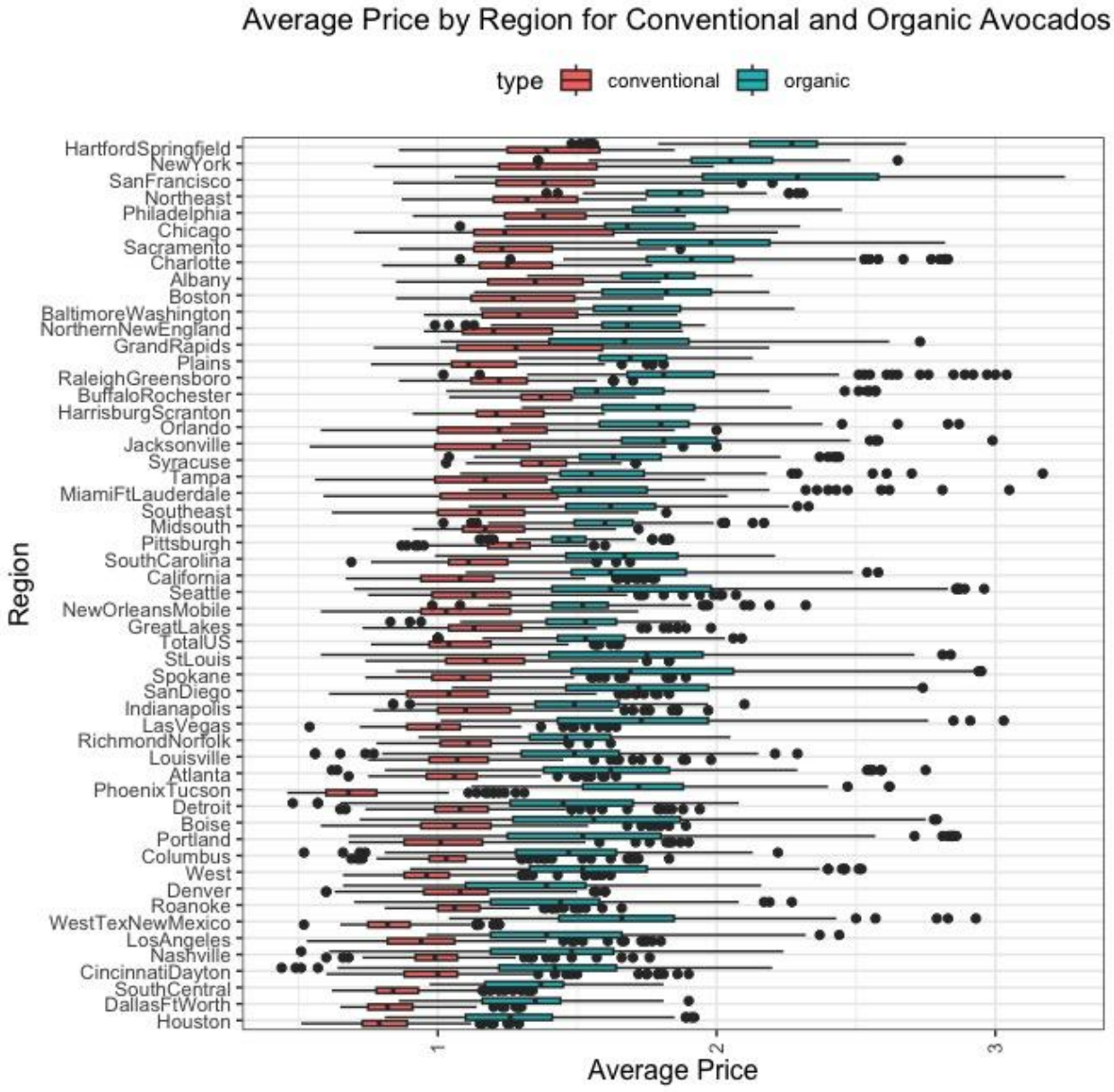


Given that avocado is a seasonal fruit, it is expected for the average price to fluctuate during the year. This plot shows that there is variation in the average price of avocado princes over the months of any year.



Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

For both conventional and organic avocados, the price slowly rises during the fall and spikes around October.



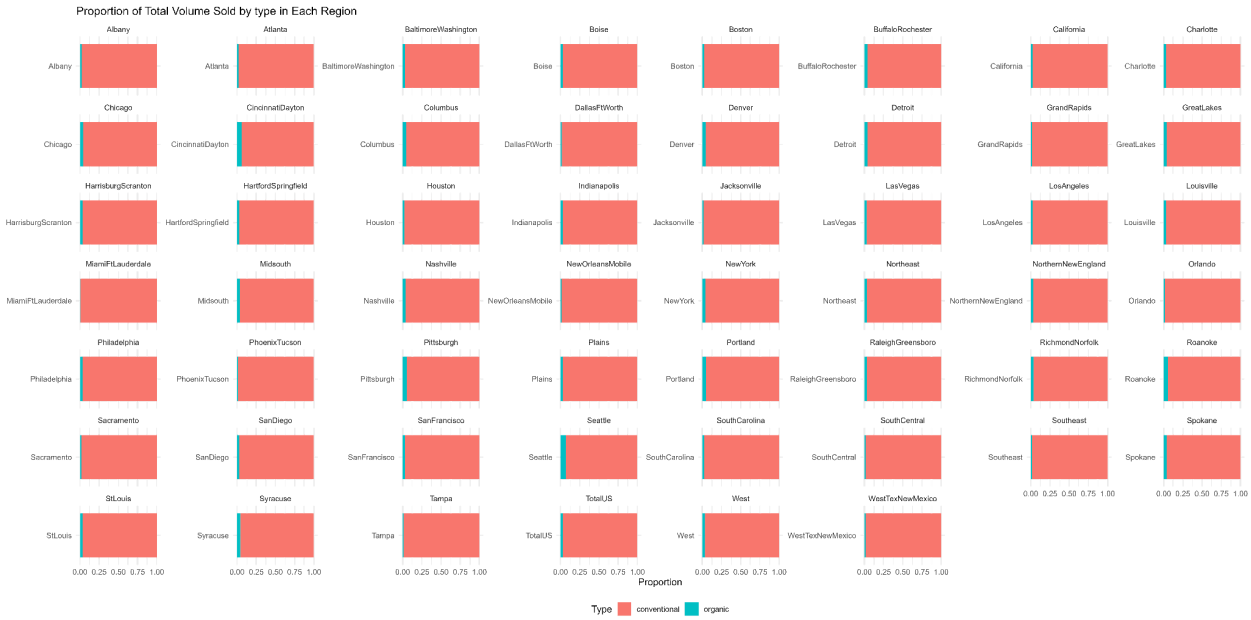
The plot above shows the average price for the two types of avocado by region. Notice that the price of organic avocados are more expensive in all regions provided in the dataset.

Sample Proportion:

First, we try to analyze the proportion of avocados sold by type (organic vs. conventional) in each region. This helps us understand how consumer preferences for avocado types vary by region.

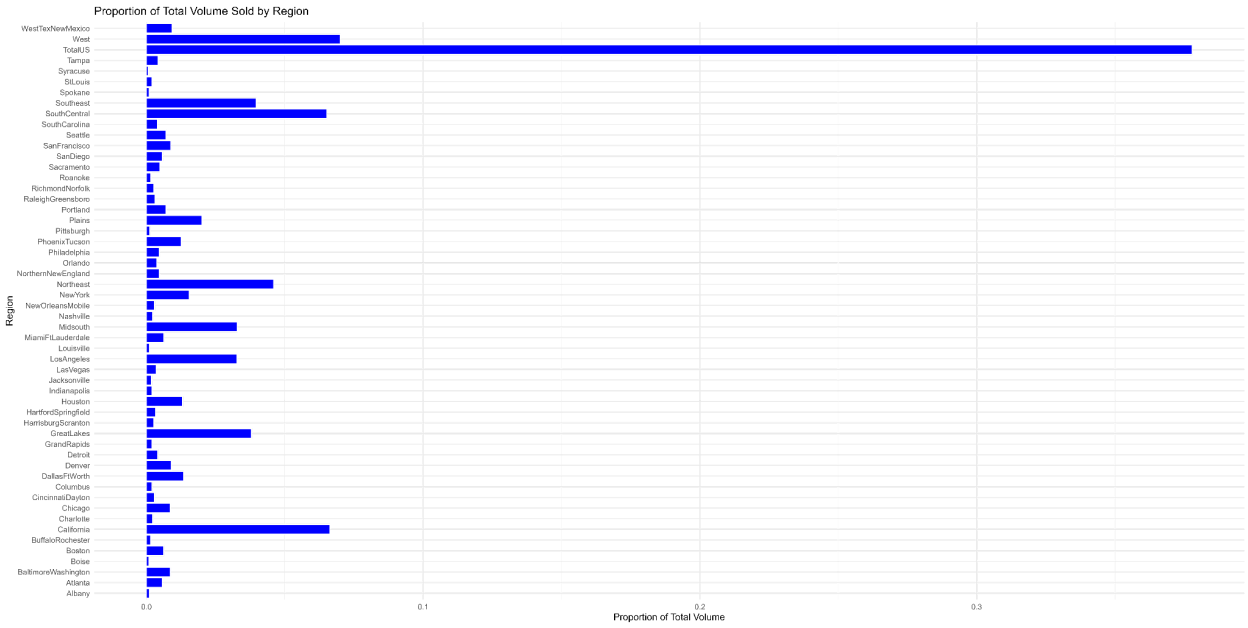
Applied Stats Project Writeup

Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



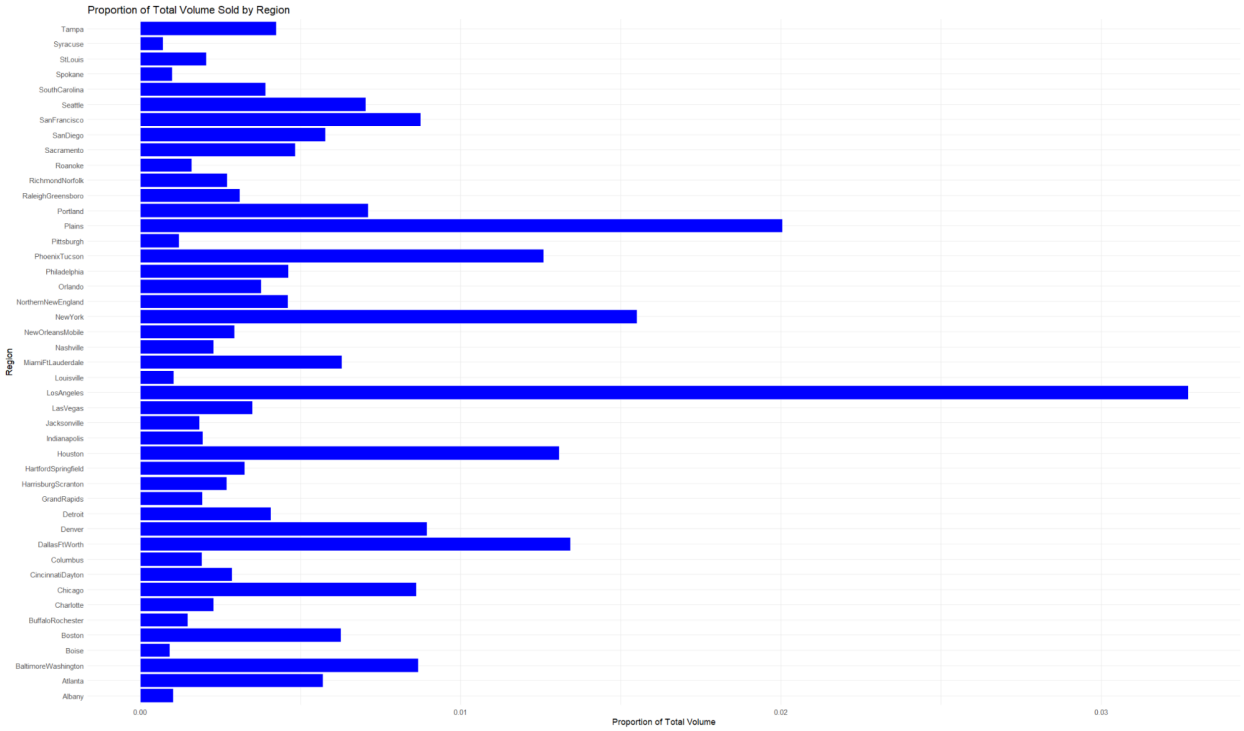
According to these histograms, the proportion of organic avocados in the avocado market is no more than 5% in each region. From the above information, organic avocados are generally more expensive than conventional avocados. This may be a factor that causes this phenomenon.

By analyzing the proportion of total volume sold in each region, we can gain insights into which regions has a higher demand for avocados.



Since the largest regions will have a strong impact on the proportion, we decided to drop the data of the largest regions: West, Northeast and TotalUS.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

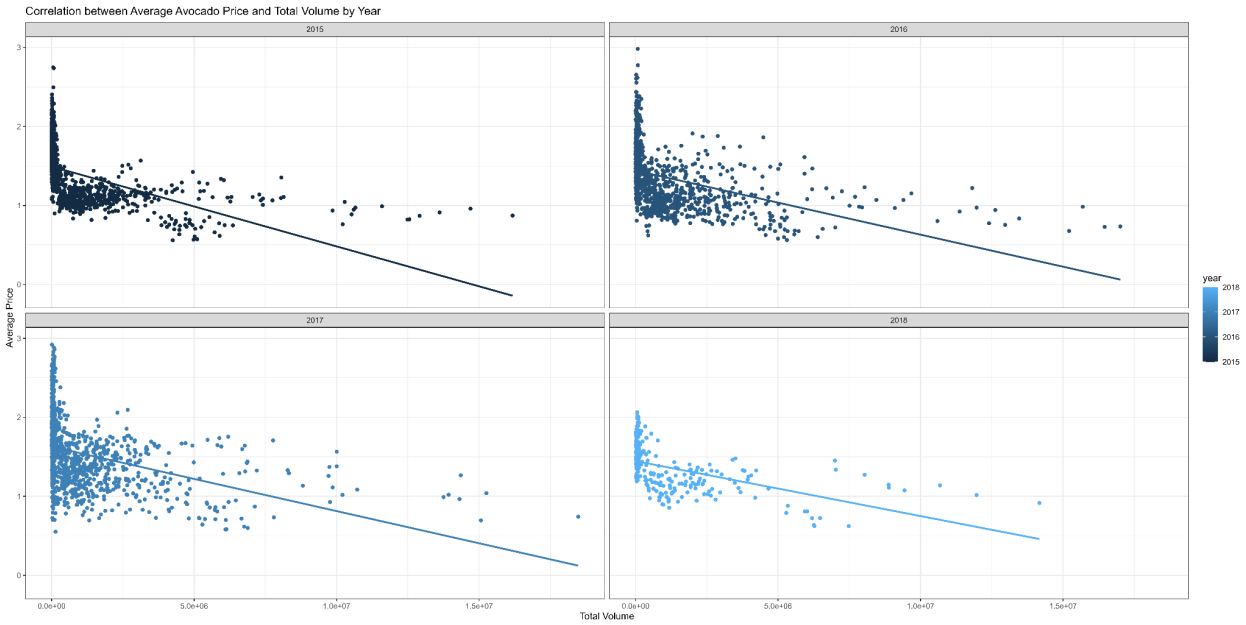


This figure shows the proportion of total volume sold by region. Plains, PhoenixTuscon, New York, Los Angeles, Houston and DallasFtWorth have a relatively higher proportion of avocado market share than other regions. Therefore, these regions may have a higher concentration of consumers who are interested in and willing to pay for avocados.

Correlation:

We also looked at whether there is a connection between the average price of avocados and how many avocados are sold each year. This might tell us if people buy more or less avocados depending on the price.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



From these figures, the correlation between total volume and average avocado price appears negative. This is in line with the basic economic theory of supply and demand. As supply decreases the price of the good will rise.

year <dbl>	correlation <dbl>
2015	-0.5103576
2016	-0.4216431
2017	-0.3852992
2018	-0.5055330

Across the board, correlation between volume and avocado price is negative. A correlation score of -0.51 between the volume of avocados sold and the average price in 2015 suggests a moderate negative relationship between these two variables. This means that as the average price of avocados increases, the volume of avocados sold tends to decrease, and vice versa. Keep in mind that there are many other factors that may cause the avocado price to fluctuate, not just the volume sold.

Correlation Test:

By conducting a correlation test, we can determine if there is a linear association between the two variables - volume and avocado price. To do so, we establish a null hypothesis that the correlation between the total volume of avocados sold and the price of avocados is equal to zero. This suggests that there is no relationship between these variables in the dataset.

```
##
##  Pearson's product-moment correlation
##
## data:  regiondata$AveragePrice and regiondata$`Total Volume`
## t = -58.283, df = 15208, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4401999 -0.4142173
## sample estimates:
##        cor
## -0.4272968
```

The p-value is less than 0.05, which indicates that this correlation is statistically significant, and we can reject the null hypothesis that the true correlation between these variables is zero.

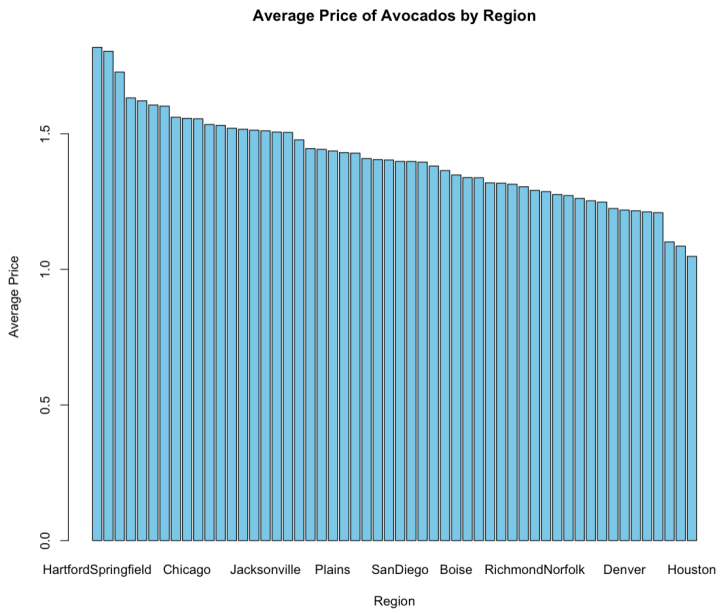
The 95% confidence interval in a correlation test represents a range of values within which we can be 95% confident that the true population correlation coefficient lies. In our case, the true correlation will lie between -0.44% and -0.41% 95% of the time.

Anova Test:

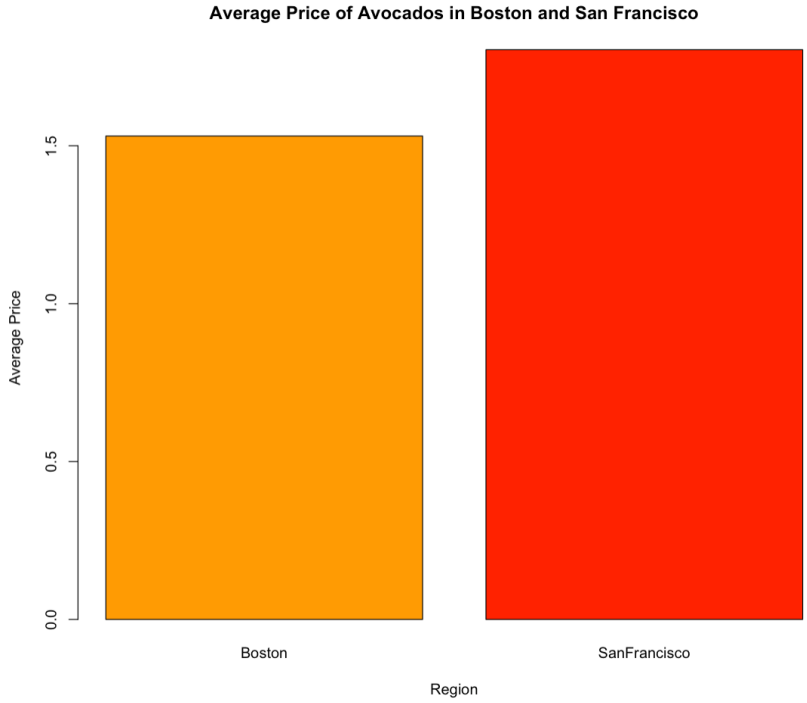
There is a region column in the dataset. We will use this column to conduct tests to see whether regions have different avocado prices. Below is a list of all the regions.

```
"Albany"              "Atlanta"             "BaltimoreWashington" "Boise"               "Boston"              "BuffaloRochester"
"California"          "Charlotte"           "Chicago"             "CincinnatiDayton"    "Columbus"            "DallasFtWorth"
"Denver"              "Detroit"             "GrandRapids"         "GreatLakes"          "HarrisburgScranton"  "HartfordSpringfield"
"Houston"             "Indianapolis"        "Jacksonville"        "LasVegas"            "LosAngeles"          "Louisville"
"MiamiFtLauderdale"   "Midsouth"            "Nashville"           "NewOrleansMobile"    "NewYork"             "Northeast"
"NorthernNewEngland"  "Orlando"             "Philadelphia"        "PhoenixTucson"       "Pittsburgh"          "Plains"
"Portland"            "RaleighGreensboro"   "RichmondNorfolk"     "Roanoke"             "Sacramento"          "SanDiego"
"SanFrancisco"        "Seattle"             "SouthCarolina"       "SouthCentral"        "Southeast"           "Spokane"
"StLouis"             "Syracuse"            "Tampa"               "TotalUS"             "West"                "WestTexNewMexico"
```

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



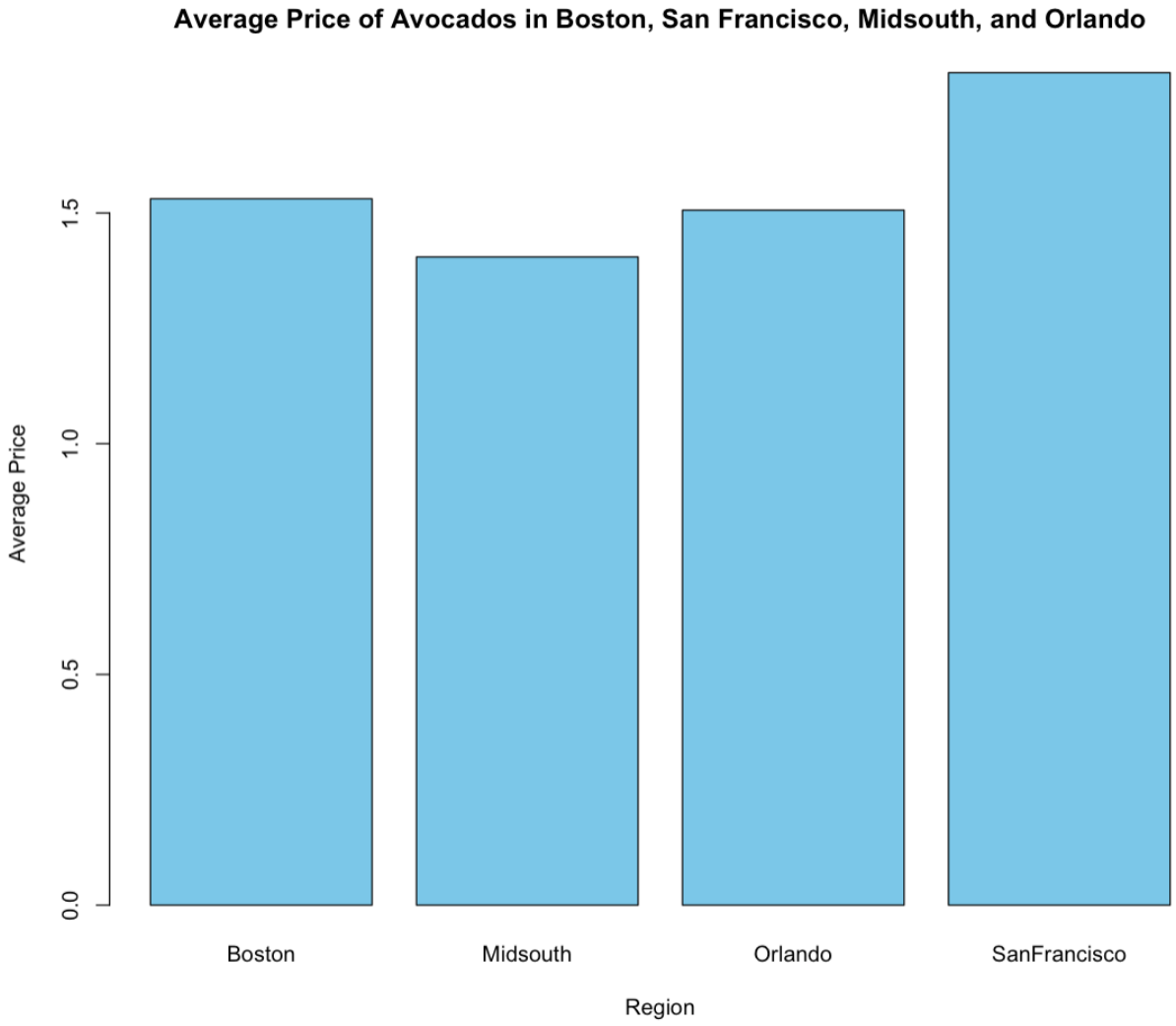
There is a fairly wide distribution of average prices between the regions in the dataset as can be seen in the graph above. The most expensive city is HartfordSpringfield with an average price of $1.82 per avocado. The cheapest city is Houston at only $1.04 per avocado. This data will allow us to use the anova test to see if there are significant differences in the average price of avocados per city.



We know that for only two samples, the anova test behaves the same as the two pair t-test. For just the city of Boston and San Francisco, we will test to see if there is a significant difference in the average avocado price.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
region        1  12.62  12.624   59.09 5.34e-14 ***
Residuals   674 143.99   0.214
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value evaluates to 5.34e-14 < 0.05. At this significance level, we reject the null hypothesis and assume that there is enough evidence to assume a difference in average avocado price between Boston and San Francisco.



Now we will test the average price between more than just two regions - Boston, Midsouth, Orlando, and San Francisco. Looking at the plot above, we can see some differences, but an anova test will statistically confirm if there are significant differences.

```
              Df Sum Sq Mean Sq F value Pr(>F)
region         3  29.56   9.854   60.52 <2e-16 ***
Residuals   1348 219.50   0.163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the result of the anova test, we can conclude that there are differences among the average avocado prices per region in at least two of the regions because the p-value is much less than 0 at 2e-16.

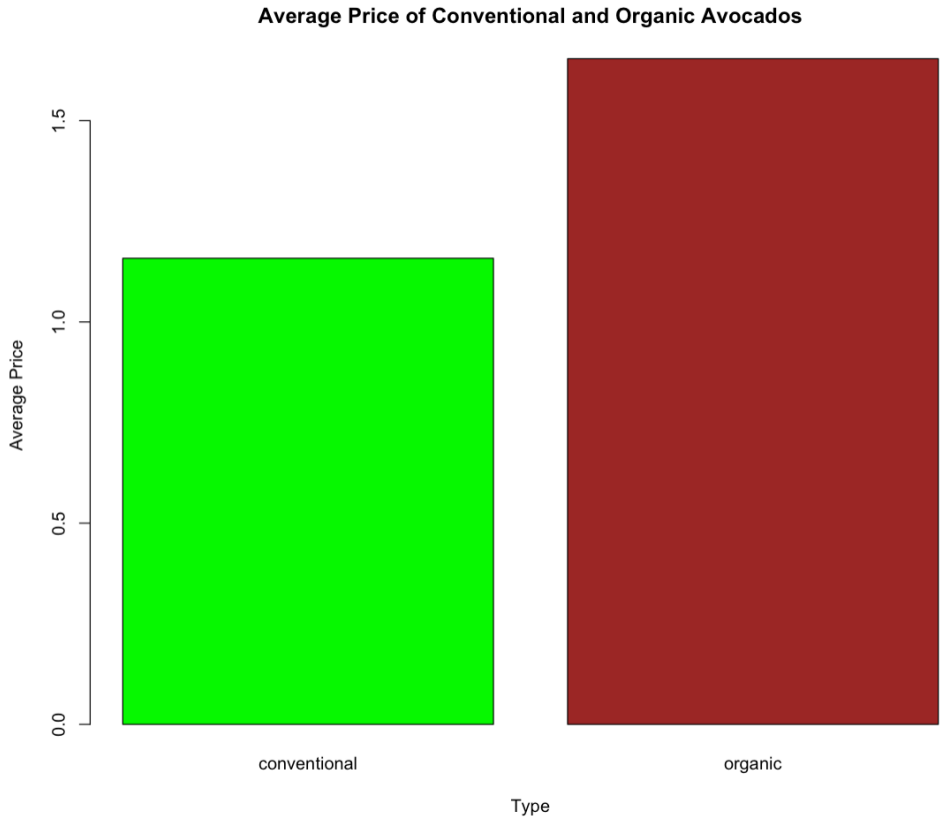
However, if you look at the graph, the difference between Orlando and Boston is only $0.025. Given that this is a small difference, we will conduct a post-hoc Tukey test to see if there are differences between that specific pair.

	diff	lwr	upr	p adj
Midsouth-Boston	-0.12612426	-0.20596868	-0.04627984	0.0002985
Orlando-Boston	-0.02467456	-0.10451897	0.05516986	0.8567781
SanFrancisco-Boston	0.27331361	0.19346919	0.35315803	0.0000000
Orlando-Midsouth	0.10144970	0.02160529	0.18129412	0.0060991
SanFrancisco-Midsouth	0.39943787	0.31959345	0.47928229	0.0000000
SanFrancisco-Orlando	0.29798817	0.21814375	0.37783258	0.0000000

The Tukey test confirms that the Orlando-Boston pair does not have a significant difference in average avocado price as the adjusted p-value is greater than 0.05 at 0.857.

There is also a "type" column that shows whether the avocado was organic or conventional. We will test to show that organic avocados and conventional avocados have different average prices.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



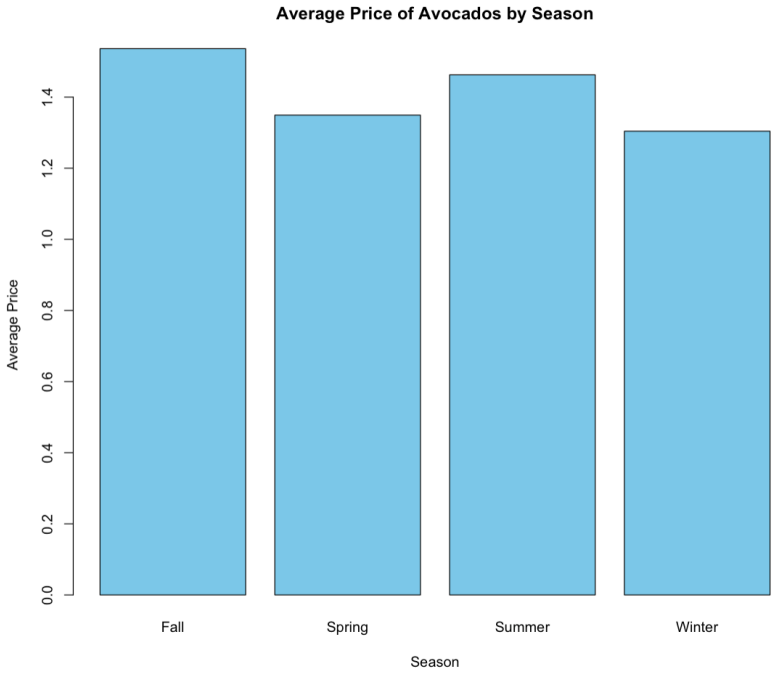
According to the plot, organic avocados on average are much more expensive.

```
              Df Sum Sq Mean Sq F value Pr(>F)
type           1   1122  1122.2   11149 <2e-16 ***
Residuals  18247   1837     0.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of 2e-16 < 0.05 confirms that the average prices of conventional avocados and organic avocados are indeed different since we reject the null hypothesis that there is no difference in the means.

Given that avocado is a seasonal fruit, we can test to see if there is a difference between mean prices per season. We use the date column in the dataframe to map each date (e.g 2022-12-01) to a season of Summer, Fall, Winter, and Spring.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



According to the plot, there appear to be differences among the average prices.

```
              Df Sum Sq Mean Sq F value Pr(>F)
Season         3  152.8   50.94   331.2 <2e-16 ***
Residuals   18245 2806.1    0.15
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test's p-value is 2e-16, which is much smaller than 0.05. Thus, we reject the null hypothesis and conclude that there is a significant difference in the average avocado price in at least two of the seasons. We will use a post-hoc Tukey test to see if there are any pairs that don't have enough evidence to prove a difference in average avocado prices.

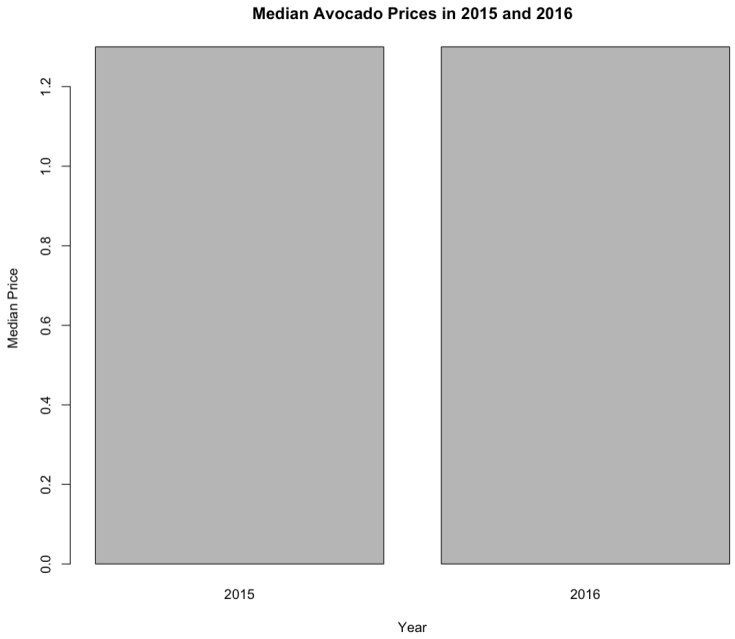
```
                    diff         lwr         upr p adj
Spring-Fall   -0.18688115 -0.20820452 -0.16555779 0e+00
Summer-Fall   -0.07357428 -0.09553313 -0.05161543 0e+00
Winter-Fall   -0.23204623 -0.25304833 -0.21104413 0e+00
Summer-Spring  0.11330687  0.09198082  0.13463292 0e+00
Winter-Spring -0.04516508 -0.06550464 -0.02482552 1e-07
Winter-Summer -0.15847195 -0.17947678 -0.13746713 0e+00
```

The Tukey test shows that all pairs have an adjusted p-value of less than 0.05, so we conclude that every season pair has a difference in their average avocado prices.

Non Parametric Tests:

While average avocado prices are normally distributed, we can still use non-parametric tests to test certain situations. We will test whether there is a statistically significant difference in median avocado prices between groups.

We will test to see if there is a difference in median price between 2015 and 2016 avocado prices. First, we filter on the date column for 2015 and 2016. Then we take 500 random samples, because for the Wilcoxon test, we need an equal number of samples per group. This will allow us to compare the median prices between the two years.



The plot shows that both medians are 1.3.

```
        Wilcoxon signed rank test with continuity correction

data:  samples$AveragePrice[samples$year == 2015] and samples$AveragePrice[samples$year == 2016]
V = 63042, p-value = 0.6017
alternative hypothesis: true location shift is not equal to 0
```

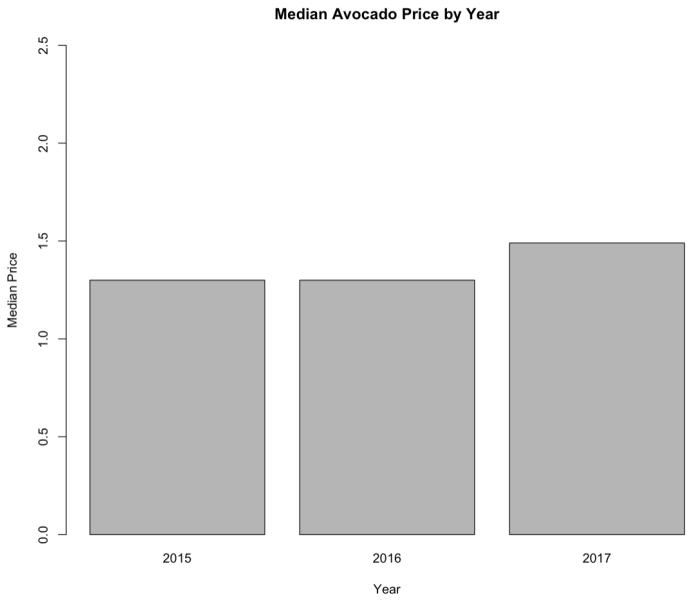
The signed rank test shows that there is no significant difference in the medians as the p-value is greater than 0.05, so we fail to reject the null hypothesis.

Based on the anova test performed earlier in the report, we saw that the average price between conventional and organic avocados are significantly different. We also performed a Wilcoxon rank sum test to confirm this observation.

```
        Wilcoxon rank sum test with continuity correction

data:  samples$AveragePrice[samples$type == "organic"] and samples$AveragePrice[samples$type == "conventional"]
W = 224777, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Since the p-value is much less than 0, we also reject the null hypothesis and confirm that the median avocado price between organic and conventional avocados are significantly different.

We also want to use the Kruskal Wallis test on the average price data for the years 2015, 2016, and 2017.



The median price for 2017 is higher than the median price of 2015 and 2016.

```
        Kruskal-Wallis rank sum test

data:  AveragePrice by year(Date)
Kruskal-Wallis chi-squared = 580.41, df = 2, p-value < 2.2e-16
```

We can conclude that there is a significant difference in the median price between the three years. However, the Kruskal-Wallis test only tells us that there is a difference between the groups, but it does not tell us which groups are different. However, based on the Wilcoxon test that we showed earlier, that showed that there was no difference in the median price between 2015 and 2017, it is fair to assume there is a difference in median avocado price between 2015-2017 and 2016-2017.

Welch two sample t-test:

The use of Welch's t-test can be used to compare the mean prices of the two avocado types. The analysis can provide a more accurate understanding of the statistical significance of the price difference between the two types of avocados.

```
        Welch Two Sample t-test

data:  conventional_data$AveragePrice and organic_data$AveragePrice
t = -105.58, df = 16619, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5051664 -0.4867517
sample estimates:
mean of x mean of y
 1.158040  1.653999
```

The t-test resulted in a p-value of less than 2.2e-16, which is an extremely small value. This indicates that the difference between the average price of conventional and organic avocados is statistically significant.

Chi-squared test

We generated a table that displays the frequency distribution of avocado types and regions in the Avocado dataset. A chi-squared test is performed to investigate whether there is a significant correlation between the two categorical variables.

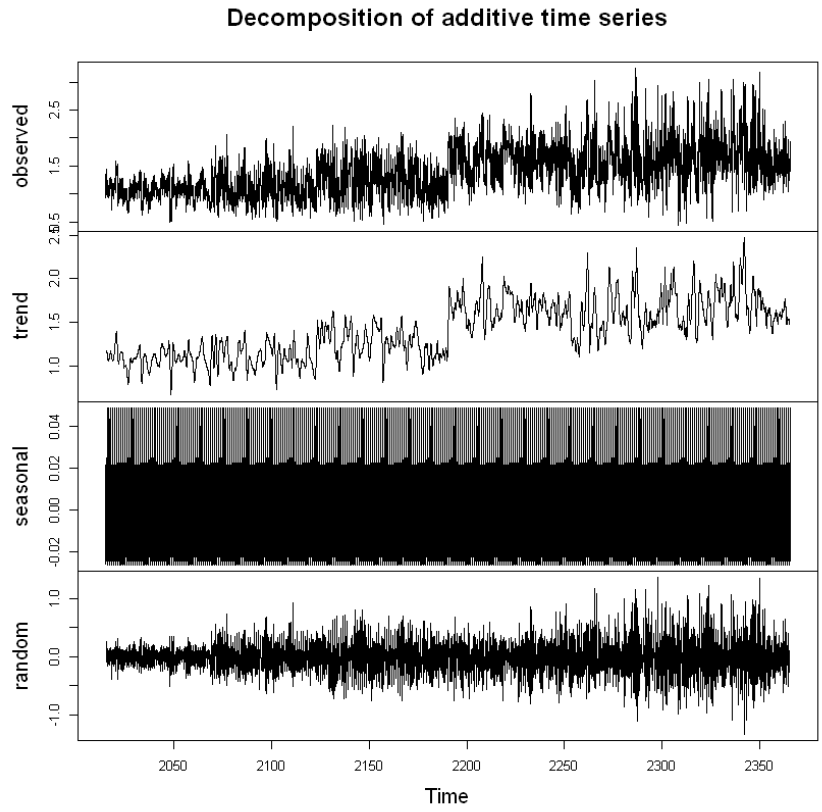
```
        Pearson's Chi-squared test

data:  cont_table
X-squared = 0.026372, df = 53, p-value = 1
```

The test statistic shows that there is no such significant correlation, as the p-value of 1 indicates that the observed distribution of avocado types and regions in the dataset is consistent with the assumption of independence between the two variables.

The chi-squared test found no such association, it can be concluded that avocado type and region are independently distributed in the dataset.

Time Series

We transformed the "Date" column in the Avocado dataset to a date format and created a weekly frequency time series of avocado prices.



The figure has four subplots:

The top subplot displays the original time series data, which is the weekly average price of avocados from 2015 to 2018.

The second subplot from the top displays the seasonal component, which represents the recurring patterns in the data that occur on a yearly basis.

The third subplot from the top shows the trend component, which represents the overall long-term behavior of the data.

The bottom subplot displays the residual component, which represents the random variation in the data that is not explained by the seasonal or trend components.

Confidence Intervals

conventional

```
Mean price: 1.15804
95 % confidence interval: [ 1.152642 , 1.163437 ]
```

We can see that the mean price of conventional avocados is 1.15804, and the 95% confidence interval is [1.152642, 1.163437]. This means that we can be 95% confident that the true population mean of the average price of conventional avocados is within this interval.

organic

```
Mean price: 1.653999
95 % confidence interval: [ 1.646539 , 1.661459 ]
```

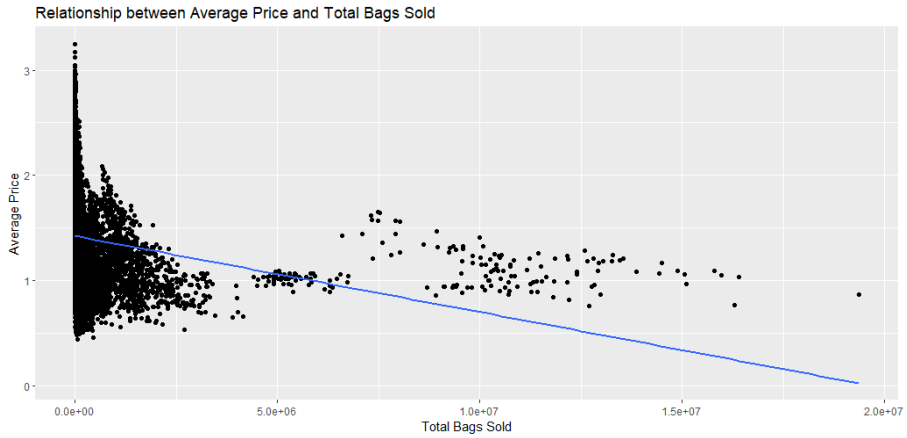
We can see that the mean price of conventional avocados is 1.653999, and the 95% confidence interval is [1.646539,1.661459]. This means that we can be 95% confident that the true population mean of the average price of organic avocados is within this interval.

Overall, organic avocados are more expensive than conventional avocados.

Simple Linear Regression

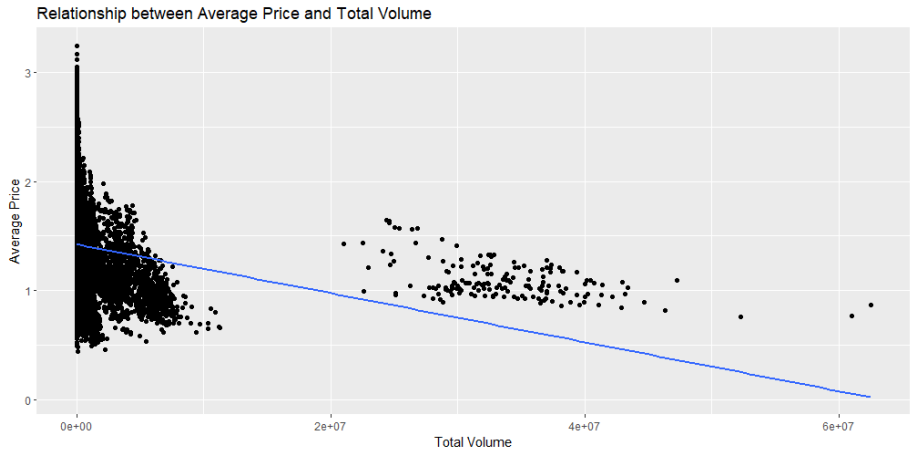
```
Call:
lm(formula = AveragePrice ~ Total.Bags, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97903 -0.30046 -0.03232  0.25174  1.82693

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.423e+00  3.019e-03   471.4   <2e-16 ***
Total.Bags  -7.230e-08  2.975e-09   -24.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3963 on 18247 degrees of freedom
Multiple R-squared:  0.03136,   Adjusted R-squared:  0.03131
F-statistic: 590.8 on 1 and 18247 DF,  p-value: < 2.2e-16
```

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying



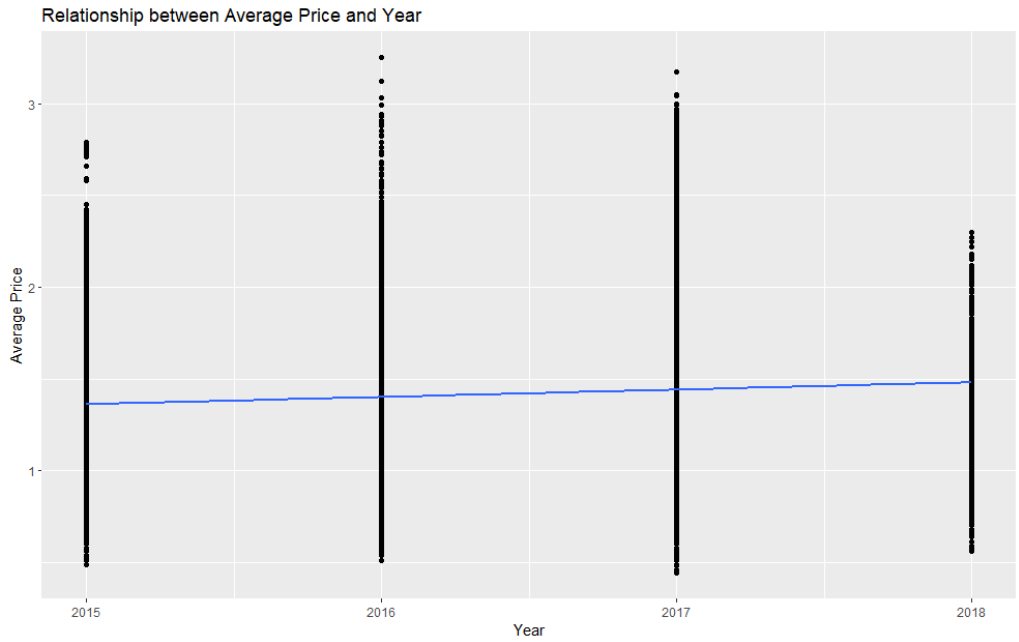
We started by comparing the average price to total bags sold to compare if there was a linear relationship present. The p-value is less than 0.05, so we can reject the null hypothesis and conclude that the coefficient is likely not equal to zero. Also, supported by the graph, there is a clear trend between average price and total bags sold. While this could be affected by several outliers the overall trend is present. More bags sold, the lower the average price.

```
Call:
lm(formula = AveragePrice ~ Total.Volume, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98366 -0.29994 -0.03319  0.25498  1.82528

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.425e+00  3.012e-03  473.07   <2e-16 ***
Total.Volume -2.247e-08  8.470e-10  -26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3951 on 18247 degrees of freedom
Multiple R-squared:  0.03715,   Adjusted R-squared:  0.0371
F-statistic: 704.1 on 1 and 18247 DF,  p-value: < 2.2e-16
```



Similarly, comparing average price to total volume shows a negative trend associated between the two variables. The small p-value shows that's a relation between the two variables and the coefficient is also likely not zero. We reject the null hypothesis and conclude the estimated coefficient of total volume is not equal to zero and is a significant predictor of average price.

```
Call:
lm(formula = AveragePrice ~ year, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00000 -0.30007 -0.04007  0.25985  1.84993

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -79.091294   6.366332  -12.42   <2e-16 ***
year          0.039926   0.003158   12.64   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4009 on 18247 degrees of freedom
Multiple R-squared:  0.008686,  Adjusted R-squared:  0.008631
F-statistic: 159.9 on 1 and 18247 DF,  p-value: < 2.2e-16
```



The linear regression formula is: price = -79.09 + year(*0.039926)
Using this formula, we can predict that 2019's avocado price will be:

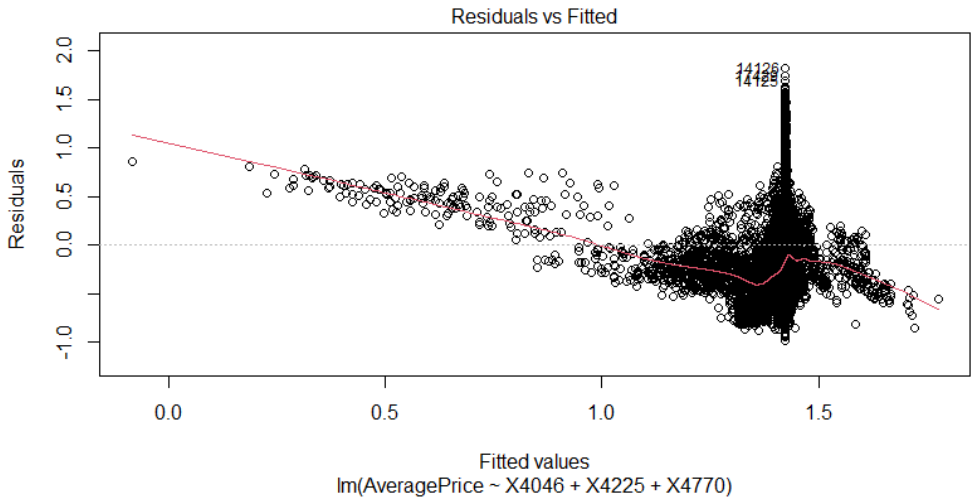
-79.09 + 2019*0.039926 = $1.521 per avocado

When comparing average price to the year, the p-value was also less than 0.05, indicating that we should reject the null hypothesis. The trend in this graph is a slight

increase over the years. We can conclude the estimated coefficient of year is not equal to zero and is a significant predictor of average price. In other words, the price of avocado has been increasing year over year and we should expect 2019's average avocado price to be larger than 2018s.

Multiple Linear Regression

```
Call:
lm(formula = AveragePrice ~ X4046 + X4225 + X4770, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98378 -0.29674 -0.03915  0.24661  1.82596

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.423e+00  2.994e-03 475.380  < 2e-16 ***
X4046       -1.051e-07  6.106e-09 -17.215  < 2e-16 ***
X4225        7.562e-08  7.705e-09   9.815  < 2e-16 ***
X4770       -3.935e-07  5.893e-08  -6.677  2.5e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3928 on 18245 degrees of freedom
Multiple R-squared:  0.04853,   Adjusted R-squared:  0.04837
F-statistic: 310.2 on 3 and 18245 DF,  p-value: < 2.2e-16
```



Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

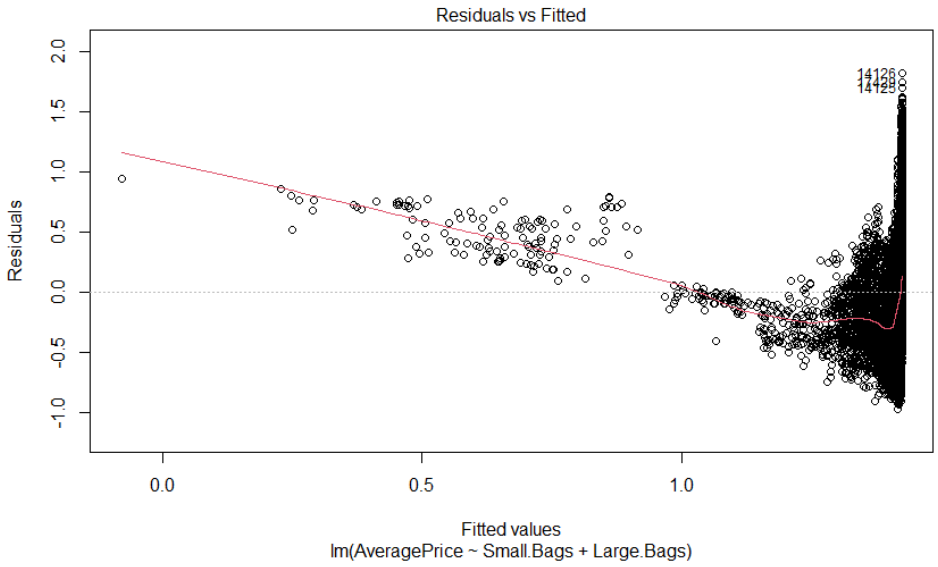
```
Analysis of Variance Table

Response: AveragePrice
             Df  Sum Sq Mean Sq F value    Pr(>F)
X4046         1  128.40 128.404 832.145 < 2.2e-16 ***
X4225         1    8.31   8.313  53.877 2.225e-13 ***
X4770         1    6.88   6.880  44.588 2.502e-11 ***
Residuals 18245 2815.29   0.154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing average price to the different sizes of avocados shows that each size the coefficient is significant given the small p-values for each. The residual error is 0.3928 which shows a low fit for the overall data. Multiple R-squared is 0.0485 meaning that only 4.8% of the variation in average price can be explained by the variation in the sizes. We can presume that there may be other important factors not considered that can help explain the variation in price. Also the f-statistic is small indicating the model is statistically significant.

Looking at the anova table provides more insight into which sizes could have significant variation in the average price. X4046 explains a substantial amount of variation on the average price as can be seen by the high F value and small p-value indicating a strong relationship between the two. X4225 and X4770 both have significant variation on the average price but when compared with X4046 they explain less variation. The residuals explain the remaining unexplained variation. It is considerably higher than the variation explained by the predictor variables meaning there are other important factors affecting avocado average prices.

```
Call:
lm(formula = AveragePrice ~ Small.Bags + Large.Bags, data = avocado)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97527 -0.30054 -0.03257  0.25247  1.82694

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.423e+00  3.019e-03 471.355  < 2e-16 ***
Small.Bags  -5.426e-08  9.131e-09  -5.943 2.86e-09 ***
Large.Bags  -1.356e-07  2.793e-08  -4.857 1.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3962 on 18246 degrees of freedom
Multiple R-squared:  0.03178,   Adjusted R-squared:  0.03168
F-statistic: 299.5 on 2 and 18246 DF,  p-value: < 2.2e-16
```

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying




```
Analysis of Variance Table

Response: AveragePrice
              Df  Sum Sq Mean Sq F value    Pr(>F)
Small.Bags     1   90.34  90.336 575.345 < 2.2e-16 ***
Large.Bags     1    3.70   3.704  23.588 1.203e-06 ***
Residuals  18246 2864.84   0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small and large bags are both negative coefficients on the average price. So as they increase, the average price decreases. They both have significant coefficients showing strong evidence in this relationship. The residual standard error is 0.39 which is a low value that should indicate a good fit but looking at the residuals vs fitted graph proves there's wide variability in the data that is also densely clustered towards one end. Only 3% of the variation in average price is explained by the bag sizes.

The ANOVA table shows that there's a considerable amount of variation in average price due the size of bags. The remaining unexplained variation in average price is explained by the residuals which is considerably higher than the variation explained by the predictor variables.

Applied Stats Project Writeup
Ido Tzhori, Noah D'Cruz, LiuYi Cui, Isaya Acevedo, KuaJiang Ying

Logistic Regression

```
Call:
glm(formula = HighPrice ~ X, family = binomial(link = "logit"),
    data = avocado)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2828  -1.1160  -0.9769   1.1949   1.4122

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.2443755  0.0276564   8.836   <2e-16 ***
X           -0.0150224  0.0009701 -15.486   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25235  on 18248  degrees of freedom
Residual deviance: 24992  on 18247  degrees of freedom
AIC: 24996

Number of Fisher Scoring iterations: 4
```

Since average price is a continuous variable. In order to perform some logistic regression on the data I converted average price to a categorical variable. Setting values above or below the mean price to either 1 or 0 respectively. We then compared the remaining features with the new price variable. The p-value is less than 0.05 so we can reject the null hypothesis. We can conclude that the features are significantly important to the average price.

```
Call:
glm(formula = HighPrice ~ year, family = binomial(link = "logit"),
    data = avocado)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.274  -1.116  -1.040   1.240   1.321

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -374.02170   32.01507  -11.68   <2e-16 ***
year           0.18545    0.01588   11.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25235  on 18248  degrees of freedom
Residual deviance: 25098  on 18247  degrees of freedom
AIC: 25102

Number of Fisher Scoring iterations: 3
```

When comparing only average price and year in a logistic regression we get the same outcome as the linear models. The factor is significantly important and the coefficient is likely not equal to zero.

Sources:

1. NeuroMUSIC. "Avocado Prices." Kaggle, 2018, https://www.kaggle.com/neuromusic/avocado-prices.

Code:

Ido Tzhori:

```
avocado <- read.csv("avocado.csv")
colnames(avocado)
unique(avocado$region)

# calculates average price by region
avg_price_by_region <- aggregate(AveragePrice ~ region, data = avocado, FUN =
mean)

# sorts the above array
avg_price_by_region_sorted <-
avg_price_by_region[order(-avg_price_by_region$AveragePrice),]

barplot(avg_price_by_region_sorted$AveragePrice, names.arg =
avg_price_by_region_sorted$region,
      main = "Average Price of Avocados by Region", xlab = "Region", ylab = "Average
Price",
      col = "skyblue")

# subset the regions
boston_sf <- subset(avocado, region %in% c("Boston", "SanFrancisco"))
avg_price_boston_sf <- aggregate(AveragePrice ~ region, data = boston_sf, FUN =
mean)

barplot(avg_price_boston_sf$AveragePrice, names.arg = avg_price_boston_sf$region,
      main = "Average Price of Avocados in Boston and San Francisco", xlab =
"Region",
      ylab = "Average Price", col = c("orange", "red"))



# anova test for just Boston and SF
avocado_bos_sf <- subset(avocado, region %in% c("Boston", "SanFrancisco"))
anova_bsf <- aov(AveragePrice ~ region, data = avocado_bos_sf)
summary(anova_bsf)

# anova test for multiple cities
bos_sf_ms_orl <- subset(avocado, region %in% c("Boston", "SanFrancisco",
"Midsouth", "Orlando"))
```

```
anova_cities <- aov(AveragePrice ~ region, data = bos_sf_ms_orl)
summary(anova_cities)

# plot of average prices
avg_price_bos_sf_ms_orl <- aggregate(AveragePrice ~ region, data = bos_sf_ms_orl,
FUN = mean)

barplot(avg_price_bos_sf_ms_orl$AveragePrice, names.arg =
avg_price_bos_sf_ms_orl$region,
      main = "Average Price of Avocados in Boston, San Francisco, Midsouth, and
Orlando",
      xlab = "Region", ylab = "Average Price", col = "skyblue")

# post hoc tukey test to compare city-city pairs
tukey_results <- TukeyHSD(anova_cities)
tukey_results

unique(avocado$type)

# average price by type
avg_price_type <- aggregate(AveragePrice ~ type, data = avocado, FUN = mean)

barplot(avg_price_type$AveragePrice, names.arg = avg_price_type$type,
      main = "Average Price of Conventional and Organic Avocados",
      xlab = "Type", ylab = "Average Price", col = c("green", "brown"))

# anova of the two types
anova_type <- aov(AveragePrice ~ type, data = avocado)
summary(anova_type)

# convert to proper date column
avocado$Date <- as.Date(avocado$Date)

avg_price_season <- aggregate(AveragePrice ~ Season, data = avocado, FUN = mean)

barplot(avg_price_season$AveragePrice, names.arg = avg_price_season$Season,
      main = "Average Price of Avocados by Season",
      xlab = "Season", ylab = "Average Price", col = "skyblue")

# calculates the season based on date roughly
```

```
avocado$Season <- factor(ifelse(months(avocado$Date) %in% c("June", "July",
"August"), "Summer",
                         ifelse(months(avocado$Date) %in% c("September", "October",
"November"), "Fall",
                                ifelse(months(avocado$Date) %in% c("December", "January",
"February"), "Winter", "Spring"))))

# anova test of the season
anova_season <- aov(AveragePrice ~ Season, data = avocado)
summary(anova_season)

# post hoc analysis of the seasonal anova
tukey_season <- TukeyHSD(anova_season)
tukey_season

install.packages("dplyr")
library(dplyr)
library(lubridate)

# just 2015 and 2016 data
subset <- avocado %>%
  filter(year(Date) %in% c(2015, 2016))

# for Wilcoxon tests we need equally sized samples
samples <- subset %>%
  group_by(year(Date)) %>%
  sample_n(500)

medians <- aggregate(AveragePrice ~ year(Date), data = subset, FUN = median)

barplot(medians$AveragePrice, names.arg = as.factor(medians$year),
        xlab = "Year", ylab = "Median Price",
        main = "Median Avocado Prices in 2015 and 2016")

# paired = TRUE, because the price of one year is not independent from the price
# of the next years
signed_rank_test <- wilcox.test(samples$AveragePrice[samples$year == 2015],
                         samples$AveragePrice[samples$year == 2016],
                         paired = TRUE)
```

```
rank_sum_test <- wilcox.test(samples$AveragePrice[samples$type == "organic"],
                             samples$AveragePrice[samples$type == "conventional"])

# multiple year subset for kw test
subset <- avocado %>%
  filter(year(Date) %in% c(2015, 2016, 2017))

kw_test <- kruskal.test(AveragePrice ~ year(Date), data = subset)

# plot the medians of avocado prices
medians <- subset %>%
  group_by(as.integer(format(Date, "%Y"))) %>%
  summarize(median_price = median(AveragePrice))

barplot(medians$median_price, names.arg = medians$`as.integer(format(Date, "%Y"))`,
        xlab = "Year", ylab = "Median Price", main = "Median Avocado Price by Year")
```

LiuYi Cui:

```
library(readr)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(lubridate)

# loading Data
avocado <- read_csv("avocado.csv")
# Split data by region and calculate proportion of total volume sold by type
avocado_prop <- avocado %>%
  group_by(region, type) %>%
  summarise(`Total Volume` = sum(`Total Volume`)) %>%
  mutate(prop = `Total Volume` / sum(`Total Volume`))
# Create the histograms to display proportion
Por1<-ggplot(avocado_prop, aes(y = region, x = prop, fill = type)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ region, scales = "free_y") +
  theme_minimal() +
  theme(legend.position = "bottom") +
  ggtitle("Proportion of Total Volume Sold by type in Each Region") +
```

```
  labs(x = "Proportion", y = NULL, fill = "Type")

#Save image to computer
ggsave(file = "avocado_hist.png", plot = Por1, width = 20, height = 10, dpi = 500)

#Split data by region and calculate the proportion of total volume by region
prop_by_all_region <- avocado %>%
  group_by(region) %>%
  summarize(prop_volume = sum(`Total Volume`)/sum(avocado$`Total Volume`))

# Create the histograms to display proportion
por3 <- ggplot(prop_by_all_region, aes(x = prop_volume, y = region)) +
  geom_histogram(stat = "identity", bins = 10, fill = "blue", color = "white") +
  ggtitle("Proportion of Total Volume Sold by Region") +
  xlab("Proportion of Total Volume") +
  ylab("Region") +
  theme_minimal()
por3

# Same as above
ggsave(file = "avocado_hist3.png", plot = por3, width = 20, height = 10, dpi = 500)

# Drop the large region data
regions_to_drop <- c("West",
"TotalUS","WestTexNewMexico","California","Northeast","Southeast","SouthCentral","Gr
eatLakes","Midsouth")
regiondata <- avocado %>% filter(!region %in% regions_to_drop)
# Recalculate the proportion
prop_by_region <- regiondata %>%
  group_by(region) %>%
  summarize(prop_volume = sum(`Total Volume`)/sum(avocado$`Total Volume`))

# Create the histogram after drop large region data
por2 <- ggplot(prop_by_region, aes(x = prop_volume, y = region)) +
  geom_histogram(stat = "identity", bins = 10, fill = "blue", color = "white") +
  ggtitle("Proportion of Total Volume Sold by Region") +
  xlab("Proportion of Total Volume") +
  ylab("Region") +
  theme_minimal()
por2
```

```
# same as above
ggsave(file = "avocado_hist2.png", plot = por2, width = 20, height = 10, dpi = 500)

#Create the two way scatter plot by total volume and averageprice each year
plt1 <- ggplot(avocado, aes(x = `Total Volume`, y = AveragePrice, color = year)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Total Volume", y = "Average Price", title = "Correlation between Average
Avocado Price and Total Volume by Year") +
  theme_bw() +
  facet_wrap(~ year, ncol = 2)

ggsave(file = "avocado_plt1.png", plot = plt1, width = 20, height = 10, dpi = 500)

# Reformat the date
str(avocado$Date)
avocado$Date <- as.Date(avocado$Date, format = "%Y-%m-%d")
regiondata$Date <- as.Date(regiondata$Date)
# Split the data by monthly average price and total volume
regiondata <- regiondata %>%
  mutate(month = floor_date(Date, unit = "month"))
monthly_data <- regiondata %>%
  group_by(region, type, month,year) %>%
  summarise(avg_price = mean(AveragePrice),
            total_volume = sum(`Total Volume`))

# Recreate the two way scatter plot by mouthly data
plt <- ggplot(monthly_data, aes(x = total_volume, y = avg_price, color = year)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Total Volume", y = "Average Price", title = "Correlation between Average
Avocado Price and Total Volume by Year") +
  theme_bw() +
  facet_wrap(~ year, ncol = 2)

# Same as above
ggsave(file = "avocado_plt.png", plot = plt, width = 20, height = 10, dpi = 500)

# Correlation between total volume and average price
```

```
avocado %>%
  group_by(year) %>%
  summarise(correlation = cor(AveragePrice, `Total Volume`))

# Correlation between same variables after drop large region data
regiondata %>%
  group_by(year) %>%
  summarise(correlation = cor(AveragePrice, `Total Volume`))

# correlation test
cor.test(x = avocado$AveragePrice, y = avocado$`Total Volume`)

# Correlation test after drop data
cor.test(regiondata$AveragePrice,regiondata$`Total Volume`)
```

Noah D'Cruz:

```
avocado <- read.csv("avocado.csv")

# conventional and organic
conventional <- subset(avocado, type == "conventional")
organic <- subset(avocado, type == "organic")

# t-test conventional and organic avocados
t.test(conventional$AveragePrice, organic$AveragePrice)


# ANOVA for avocado prices across regions
avocado_anova <- aov(AveragePrice ~ region, data = avocado)
summary(avocado_anova)

# correlation avocado prices, total volume sold
cor.test(avocado$AveragePrice, avocado$Total.Volume)

#contingency avocado type, region table
cont_table <- table(avocado$type, avocado$region)
#chi-squared test for avocado type and region
chisq.test(cont_table)

# Convert the Date column to a date format
```

```
avocado$Date <- as.Date(avocado$Date, "%Y-%m-%d")
# time series  avocado prices
avocado_ts <- ts(avocado$AveragePrice, start = c(2015, 1), frequency = 52)
# decomposition of the time series
avocado_decomp <- decompose(avocado_ts)
# Plot
plot(avocado_decomp)

# Select only the conventional avocados
conventional <- subset(avocado, type == "conventional")

# Calculate the mean and standard deviation of the average price
mean_price <- mean(conventional$AveragePrice)
sd_price <- sd(conventional$AveragePrice)
conf_level <- 0.95
n <- length(conventional$AveragePrice)
stderr <- sd_price / sqrt(n)
margin_error <- qt(conf_level/2 + 0.5, n-1) * stderr
lower_ci <- mean_price - margin_error
upper_ci <- mean_price + margin_error

# Print the results
cat("Mean price:", mean_price, "\n")
cat(conf_level * 100, "% confidence interval: [", lower_ci, ",", upper_ci, "]")

# Select only the organic avocados
organic <- subset(avocado, type == "organic")

# Calculate the mean and standard deviation of the average price
mean_price <- mean(organic$AveragePrice)
sd_price <- sd(organic$AveragePrice)
conf_level <- 0.95
n <- length(organic$AveragePrice)
stderr <- sd_price / sqrt(n)
margin_error <- qt(conf_level/2 + 0.5, n-1) * stderr
lower_ci <- mean_price - margin_error
upper_ci <- mean_price + margin_error

# Print the results
cat("Mean price:", mean_price, "\n")
```

```
cat(conf_level * 100, "% confidence interval: [", lower_ci, ",", upper_ci, "]")
```

Isaya Acevedo:

```
avocado <- read.csv(file.choose(), header = TRUE)


###LINEAR MODELS
#Linear model comparing avg price to total bags sold
lm_model <- lm(AveragePrice ~ Total.Bags, data = avocado)
summary(lm_model)

ggplot(avocado, aes(x = Total.Bags, y = AveragePrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Average Price and Total Bags Sold",
       x = "Total Bags Sold", y = "Average Price")


#Linear model comparing avg price to total volume sold
lm_model2 <- lm(AveragePrice ~ Total.Volume, data = avocado)
summary(lm_model2)

ggplot(avocado, aes(x = Total.Volume, y = AveragePrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Average Price and Total Volume",
       x = "Total Volume", y = "Average Price")

#Linear model to compare avg price and year
model_year <- lm(AveragePrice ~ year, data = avocado)
summary(model_year)

ggplot(avocado, aes(x = year, y = AveragePrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Average Price and Year",
```

```
      x = "Year", y = "Average Price")
# Subsetting data by type
avocado_organic <- subset(avocado, type == "organic")
avocado_conventional <- subset(avocado, type == "conventional")

# Linear models comparing avg price by year sorted by type of avocado
model_organic <- lm(AveragePrice ~ year, data = avocado_organic)
model_conventional <- lm(AveragePrice ~ year, data = avocado_conventional)

summary(model_organic)
summary(model_conventional)

# Graphs both linear models for conventional and organic avocados
ggplot(avocado, aes(x = year, y = AveragePrice, color = type)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Average Price and Year",
      x = "Year", y = "Average Price") +
  facet_wrap(~ type)

###MULTIPLE LINEAR MODELS

# Linear model comparing avg price to different avocado sizes
model <- lm(AveragePrice ~ X4046 + X4225 + X4770, data = avocado)
summary(model)

#Plot the residual vs fitted graph
plot(model, which = 1)
anova(model)

# Linear model comparing avg price to different avocado sizes
model3 <- lm(AveragePrice ~ Small.Bags + Large.Bags , data = avocado)
summary(model3)

# Plot the residual vs fitted graph
plot(model3, which = 1)
anova(model3)

###LOGISTIC MODELS
# Converting avg price from continuous to categorical
```

```
average_price <- mean(avocado$AveragePrice)
avocado$HighPrice <- ifelse(avocado$AveragePrice > average_price, 1, 0)


# Logistic model comparing high price to all remaining features
model <- glm(HighPrice ~ Total.Volume + X4046 + X4225 + X4770 + Total.Bags +
Small.Bags + Large.Bags + XLarge.Bags, family = binomial(link = "logit"), data =
avocado)
summary (model)


# Logistic model comparing high price to year
model <- glm(HighPrice ~ year, family = binomial(link = "logit"), data = avocado)
summary (model)
```

Kuajing Ying:

```
library(dplyr)
library(ggplot2)
library(ggtext)
library(lubridate)
library(tidyr)

rm(list=ls())

# 0: read in the dataset
avocado <- read.csv("avocado.csv", header = TRUE)

# 1: Average price based on type: Conventional vs. Organic

# 1.1 draw histograms for each type
ggplot(avocado, aes(x = AveragePrice, fill = type)) +
  geom_histogram(bins = 30, alpha = 0.8, position = 'identity', color = "black") +
  facet_wrap(~ type) +
  labs(title = "Histogram of Average Price by Type",
       x = "Average Price",
       y = "Frequency") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
```

```
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    strip.text = element_text(face = "bold", size = 12),
    legend.position = "none"
  ) +
  scale_fill_manual(values = c("Conventional" = "#56B4E9", "Organic" = "#D55E00"))


# 1.2 mean price for each type
mean_price_by_type <- aggregate(AveragePrice ~ type, data = avocado, FUN = mean)

print(mean_price_by_type)

summary_by_type <- by(avocado$AveragePrice, avocado$type, summary)

print(summary_by_type)


# 1.3 box plot
ggplot(avocado, aes(x = type, y = AveragePrice)) +
  geom_boxplot(fill = "blue", alpha = 0.5) +
  ggtitle("Average Avocado Prices by Type")



# 2: check seasonal fluctuations

# 2.1 Extract the month from the date column
avocado$month <- month(ymd(avocado$Date))

# 2.2 Summarize the average price for each month and type
monthly_summary <- avocado %>%
  group_by(type, month) %>%
  summarise(avg_price = mean(AveragePrice, na.rm = TRUE)) %>%
  arrange(type, month)

print(monthly_summary)

# 2.3 Create a line graph for monthly price
```

```
ggplot(monthly_summary, aes(x = month, y = avg_price, group = type, color = type)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = seq(1, 12, 1)) +
  labs(title = "Average Price of Avocados by Type",
       x = "Month",
       y = "Average Price",
          color = "Type") +
  theme_minimal()

# 2.4 Draw monthly price table
monthly_summary_table <- monthly_summary %>%
  pivot_wider(names_from = month, values_from = avg_price)

monthly_summary_long <- monthly_summary_table %>%
  mutate(type = as.factor(type)) %>%
  pivot_longer(-type, names_to = "month", values_to = "avg_price") %>%
  mutate(month = as.numeric(month))
ggplot(monthly_summary_long, aes(x = month, y = type, fill = avg_price, label =
round(avg_price, 2))) +
  geom_tile(color = "white") +
  geom_text(color = "white", size = 4) +
  scale_x_continuous(breaks = 1:12) +
  scale_fill_gradient2(low = "lightblue", mid = "blue", high = "darkblue", midpoint =
median(monthly_summary_long$avg_price)) +
  labs(x = "Month", y = "Type", fill = "Average Price", title = "Average Price by Month and
Type") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10),
        axis.title = element_text(size = 12),
        plot.title = element_text(size = 14, hjust = 0.5))

# 3: check geographical variations

# 3.1 Average Price by Region, conventional and organic
conventional_data <- avocado %>% filter(type == "conventional")
organic_data <- avocado %>% filter(type == "organic")

# 3.2 Plot, order cities by price from high to low
```

```
ggplot(avocado, aes(x = reorder(region, AveragePrice, median), y = AveragePrice, fill =
type)) +
  geom_boxplot(show.legend = TRUE) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.title = element_text(size = 12),
        legend.position = "top") +
  labs(x = "Region", y = "Average Price", title = "Average Price by Region for
Conventional and Organic Avocados") +
  coord_flip()
```