Project 2

Nancie Kung, Calvin Raab, David Collier and Eitan Shimonovitz

6/02/2021

Introduction

In this project, we will be analyzing Hass Avocados. Our objective is to use a time series to explain how the average price and total volume of Hass avocados have changed over time. First, we explored the data by looking at measures such as the changes in price and volume and the popularity of avocados by region. We then created an AR and ARDL model to examine the patterns in price and volume in the past. We also used these models to make predictions about how price and volume would change in the future. Below, we show our analysis.

Description of the Data

Our project uses historical data on avocado prices and sales volume in U.S. markets between the years 2015 and 2018. This data is based on the weekly retail sales of Hass avocados reported by retailers' cash registers. It contains an aggregation of data from multiple locations across the United States and multiple types of retail outlets. The average price is the per unit cost for each avocado and the Product Lookup code indicates the total number sold for a given type of Hass avocado. The following variables are included in the data set:

- Date - the date of the observation
- AveragePrice - the average price of a single avocado
- Type - conventional or organic
- Year - the year
- Region - the city or region of the observation
- Total Volume - total number of avocados sold
- Hass.Small - total number of avocados with PLU 4046 sold (Small Hass Avocados)
- Hass.Large - total number of avocados with PLU 4225 sold (Large Hass Avocados)
- Hass.Extra.Large - total number of avocados with PLU 4770 sold (Extra Large Hass Avocados)

Load Data

```
avocados <- read.csv("avocado.csv")
attach(avocados)
avocados <- avocados %>%
  rename(
    Hass.Small = X4046,
    Hass.Large = X4225,
    Hass.Extra.Large = X4770
  )
```

Creating A Time Series

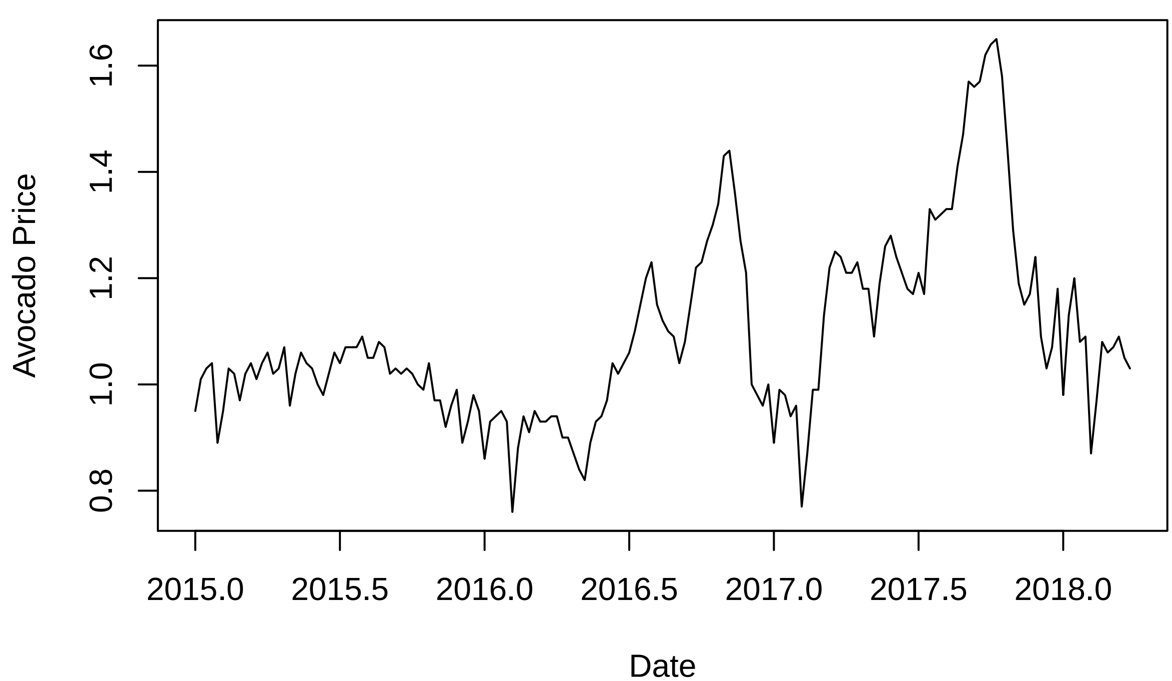
```
library(dplyr)

avocados <- avocados %>% arrange(region, type, Date)

avocados_us_conventional <- avocados %>% dplyr::filter(region == "TotalUS", type == "conventional")

avocados_us_conv.ts <- ts(avocados_us_conventional, frequency = 52, start = c(2015,1), end = c(2018,13))

avocados_us_organic <- avocados %>% dplyr::filter(region == "TotalUS", type == "organic")
avocados_us_org.ts <- ts(avocados_us_organic, frequency = 52, start = c(2015,1), end = c(2018,13))
```

Exploratory Analysis

Average Price of Avocados Over Time

From the graph below it can be seen that prices appear to spike around the early summer months, right prior to the halfway point in the year. This means that avocados are at their peak price right around now.

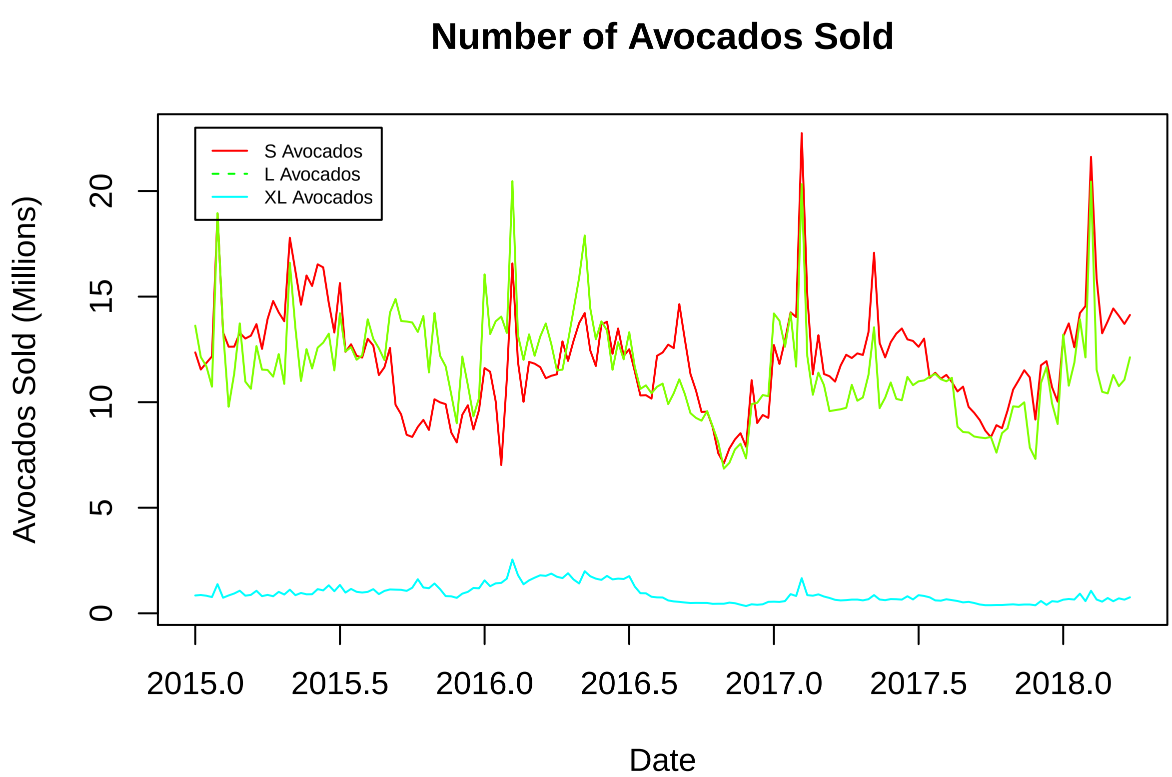
```
plot(avocados_us_conv.ts[,3], ylab = "Avocado Price", xlab = "Date")
```



Difference In Number of Avocados Sold Overtime for Different Size Haas Avocados

From the graph below it can be seen that small and large avocados appear to track one another closely and are relatively close in number of avocados sold. This graph also demonstrates that XL avocados do not sell nearly as many as small and large avocados. Small and large avocados also appear to spike around the same time.

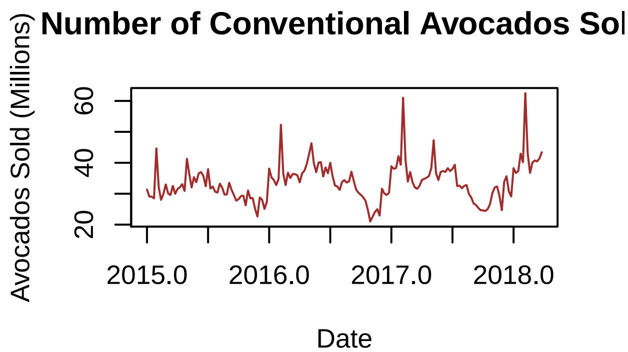
```
ts.plot(avocados_us_conv.ts[,5:7]/1000000, gpars=list(col=rainbow(4)),ylab = "Avocados Sold (Millions)"
legend(2015, 23, legend=c("S Avocados", "L Avocados", "XL Avocados"),
       col=c("red", "green", "cyan"), lty=1:2, cex=0.6)
```

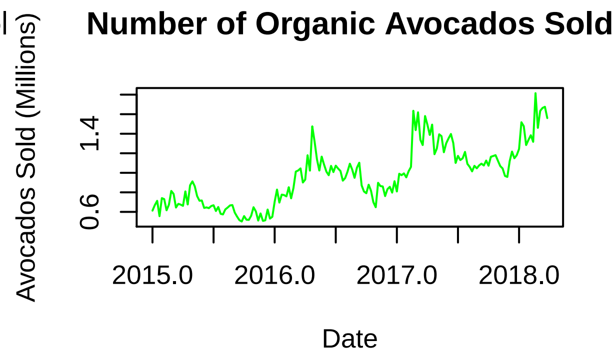


Difference In Number of Avocados Sold Over Time for Conventional vs Organic Avocados

From the graph below it can be seen that avocado sails have increased throughout the years.

```
par(mfrow=c(2,2))
plot((avocados_us_conv.ts[,4] / 1000000), ylab = "Avocados Sold (Millions)", xlab = "Date", main = "Numl
plot((avocados_us_org.ts[,4] / 1000000), ylab = "Avocados Sold (Millions)", xlab = "Date", main = "Numbe
```





Conventional avocados show strong seasonality, with spikes in the beginning of the year, while organic avocados show a consistent upwards trend from 2015 - 2018. Organic avocados appear to show their seasonality through severe dips at the end of the year, in the beginning of winter.

Average Price By Region

According to the data if you wish to by avocados at the cheapest price, you should go to Houston.

```r
avg_price_by_region <- aggregate(x = avocados$AveragePrice,
          by = list(avocados$region),
          FUN = mean)

#Top 10 Most Expensive Regions
slice(arrange(avg_price_by_region, desc(x)), 1:10)
```

```
##                Group.1        x
## 1  HartfordSpringfield 1.818639
## 2        SanFrancisco 1.804201
## 3             NewYork 1.727574
## 4        Philadelphia 1.632130
## 5          Sacramento 1.621568
## 6           Charlotte 1.606036
## 7           Northeast 1.601923
## 8              Albany 1.561036
## 9             Chicago 1.556775
## 10   RaleighGreensboro 1.555118
```

```r
#Top 10 Least Expensive Regions
slice(arrange(avg_price_by_region, x), 1:10)
```

```
##             Group.1        x
## 1           Houston 1.047929
## 2      DallasFtWorth 1.085592
## 3       SouthCentral 1.101243
## 4  CincinnatiDayton 1.209201
## 5         Nashville 1.212101
## 6        LosAngeles 1.216006
## 7            Denver 1.218580
## 8      PhoenixTucson 1.224438
## 9           Roanoke 1.247929
## 10         Columbus 1.252781
```

Data Analysis/Model

```r
y <- avocados_us_conv.ts[, "Total.Bags"]
ar_mod1 <- ar(y, aic = FALSE, order.max=2, method="ols")
summary(ar_mod1)
```

```
##             Length Class  Mode
## order         1    -none- numeric
## ar            2    -none- numeric
## var.pred      1    -none- numeric
## x.mean        1    -none- numeric
## x.intercept   1    -none- numeric
## aic           1    -none- numeric
## n.used        1    -none- numeric
## n.obs         1    -none- numeric
## order.max     1    -none- numeric
## partialacf    0    -none- NULL
## resid       169    ts     numeric
## method        1    -none- character
## series        1    -none- character
## frequency     1    -none- numeric
## call          5    -none- call
## asy.se.coef   2    -none- list
```

```r
ar_mod1
```

```
##
## Call:
## ar(x = y, aic = FALSE, order.max = 2, method = "ols")
##
## Coefficients:
##      1       2
## 0.7153  0.2313
##
```

```
## Intercept: 86211 (93883)
##
## Order selected 2  sigma^2 estimated as  1.467e+12
```

```r
forecast(ar_mod1, 52)
```

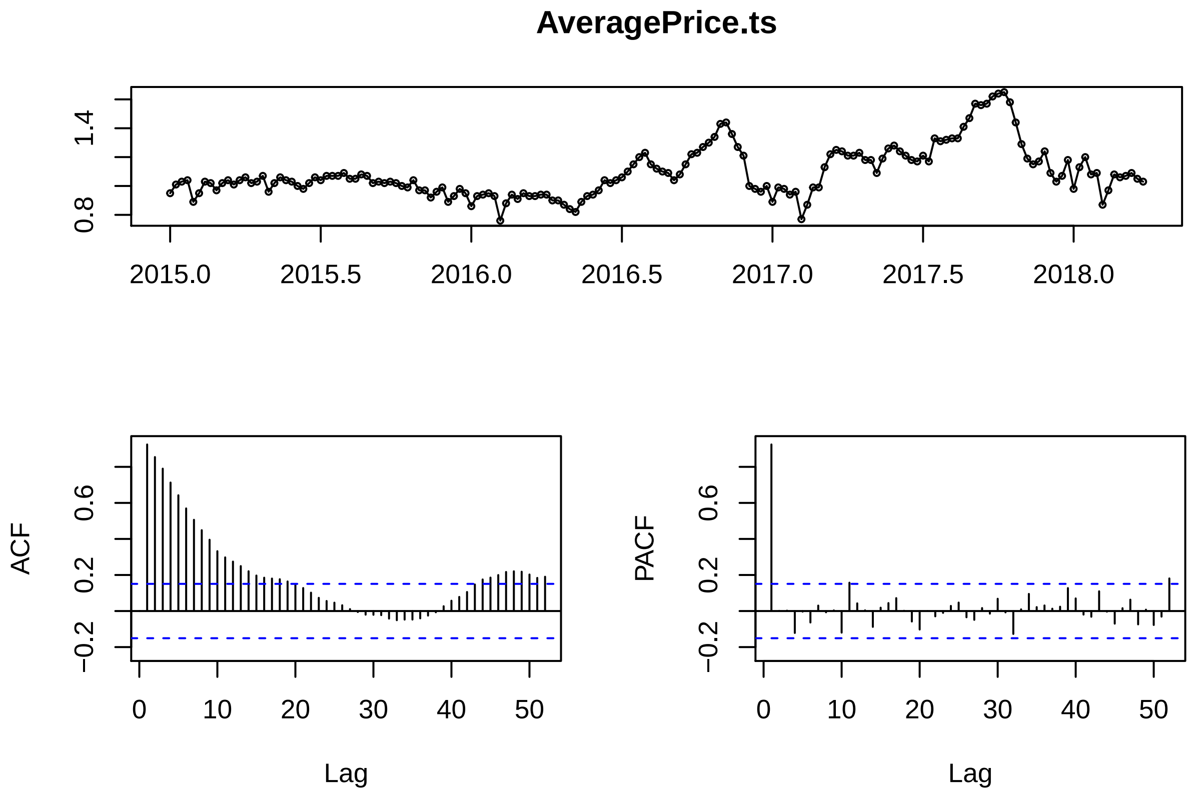
```
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2018.250       15998434 14446139 17550729 13624404 18372464
## 2018.269       15812757 13904251 17721262 12893949 18731564
## 2018.288       15588317 13358426 17818209 12177992 18998642
## 2018.308       15384829 12906403 17863255 11594404 19175254
## 2018.327       15187360 12499368 17875352 11076431 19298288
## 2018.346       14999042 12132975 17865109 10615771 19382314
## 2018.365       14818663 11798607 17838719 10199886 19437440
## 2018.385       14646079 11491602 17800556  9821723 19470435
## 2018.404       14480907 11208156 17753657  9475667 19486147
## 2018.423       14322840 10945402 17700278  9157494 19488186
## 2018.442       14171569 10701022 17642116  8863826 19479312
## 2018.462       14026804 10473123 17580484  8591918 19461689
## 2018.481       13888263 10260116 17516410  8339491 19437035
## 2018.500       13755680 10060649 17450711  8104618 19406742
## 2018.519       13628798  9873557 17384040  7885652 19371944
## 2018.538       13507373  9697823 17316922  7681169 19333576
## 2018.558       13391168  9532551 17249786  7489922 19292415
## 2018.577       13279961  9376946 17182976  7310815 19249107
## 2018.596       13173536  9230298 17116774  7142875 19204198
## 2018.615       13071687  9091969 17051406  6985233 19158142
## 2018.635       12974218  8961379 16987057  6837111 19111326
## 2018.654       12880941  8838007 16923874  6697807 19064074
## 2018.673       12791674  8721373 16861975  6566686 19016662
## 2018.692       12706246  8611041 16801450  6443171 18969321
## 2018.712       12624491  8506611 16742371  6326737 18922246
## 2018.731       12546252  8407714 16684791  6216903 18875601
## 2018.750       12471378  8314009 16628746  6113231 18829524
## 2018.769       12399723  8225183 16574262  6015316 18784130
## 2018.788       12331149  8140946 16521353  5922786 18739513
## 2018.808       12265524  8061026 16470023  5835299 18695750
## 2018.827       12202722  7985174 16420269  5752539 18652904
## 2018.846       12142619  7913156 16372083  5674213 18611025
## 2018.865       12085101  7844754 16325449  5600050 18570153
## 2018.885       12030057  7779767 16280347  5529799 18530315
## 2018.904       11977380  7718003 16236756  5463225 18491534
## 2018.923       11926967  7659286 16194649  5400112 18453822
## 2018.942       11878723  7603450 16153996  5340258 18417188
## 2018.962       11832553  7550340 16114767  5283473 18381634
## 2018.981       11788369  7499808 16076929  5229582 18347156
## 2019.000       11746084  7451719 16040449  5178420 18313749
## 2019.019       11705618  7405944 16005292  5129834 18281402
## 2019.038       11666892  7362361 15971423  5083680 18250104
## 2019.058       11629831  7320857 15938805  5039824 18219838
## 2019.077       11594364  7281325 15907403  4998140 18190589
## 2019.096       11560422  7243663 15877181  4958509 18162336
## 2019.115       11527940  7207777 15848103  4920820 18135059
## 2019.135       11496854  7173576 15820132  4884971 18108738
```

```
## 2019.154       11467105  7140976 15793234  4850861 18083349
## 2019.173       11438635  7109897 15767374  4818401 18058870
## 2019.192       11411390  7080263 15742517  4787503 18035277
## 2019.212       11385316  7052003 15718630  4758085 18012547
## 2019.231       11360364  7025049 15695678  4730072 17990656
```

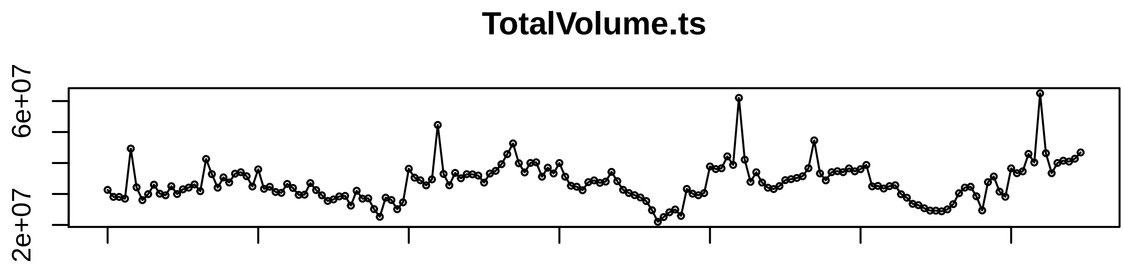
AR Process

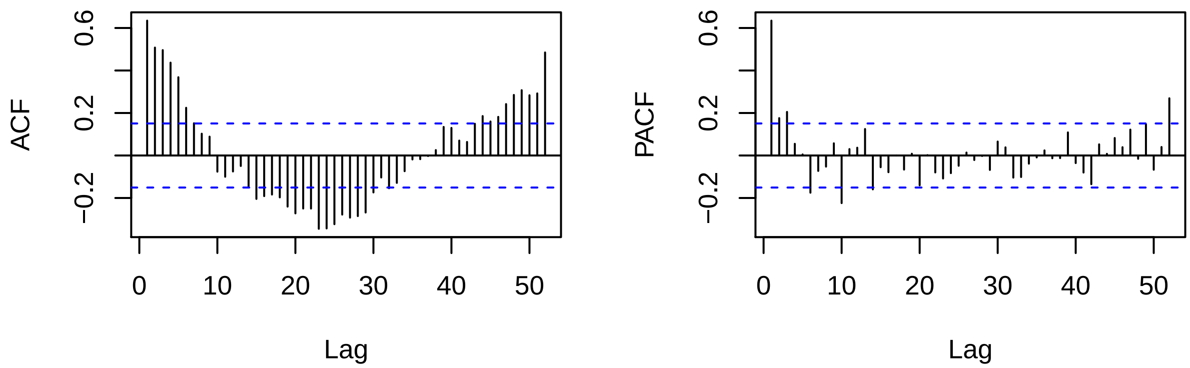
In order to see if this avocado data is cyclical, we will look at the ACF of both the Average Price and the Total Volume of our avocados. We set the lag to be maxed at 52, because there are 52 weeks in a year. From the ACF is can be seen that in the middle of the year there isn't much correlation, however near the end of the year we start to see significant statistical correlation. This tells us that there appears to be a cyclical, yearly relationship between our data. This means that data from 12 months ago can help predict the data of today.

```r
library(tseries)
library(forecast)
AveragePrice.ts <- avocados_us_conv.ts[,3]
TotalVolume.ts <- avocados_us_conv.ts[,4]
tsdisplay(AveragePrice.ts, lag.max = 52)
```




```r
tsdisplay(TotalVolume.ts, lag.max = 52)
```





The pattern of a steadily decreasing ACF with a single spike at lag = 1 for the PACF shows the pattern of an AR(1) process for average price. In addition, the pattern of a steadily decreasing ACF with multiple spikes at lag = 1,2,3,13,14 suggest a higher order AR process for average volume.

Prediction with our AR model

Here is an AR prediction model. We built a model to predict average price. The data we fed into our model was subsetted so we could use the final 5 results to test how accurate our model is. From the results below it can be seen that our prediction model did a good job and all of our confidence intervals created by the AR model contained the actual prices.

```
avo.ar <- ar(AveragePrice.ts[1:164], aic = FALSE, order.max=1, method = "ols")
summary(avo.ar)
```

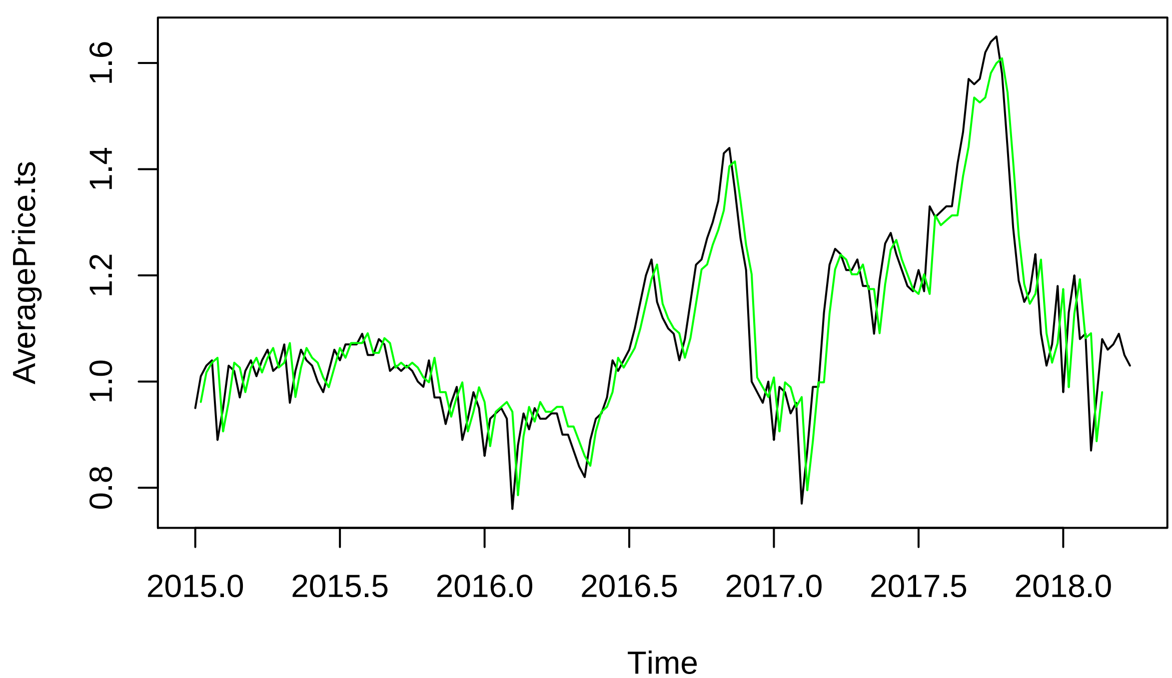
```
##              Length Class  Mode
## order          1    -none- numeric
## ar             1    -none- numeric
## var.pred       1    -none- numeric
## x.mean         1    -none- numeric
## x.intercept    1    -none- numeric
## aic            1    -none- numeric
## n.used         1    -none- numeric
## n.obs          1    -none- numeric
## order.max      1    -none- numeric
## partialacf     0    -none- NULL
## resid        164    -none- numeric
```

```
## method          1     -none- character
## series          1     -none- character
## frequency       1     -none- numeric
## call            5     -none- call
## asy.se.coef     2     -none- list
```

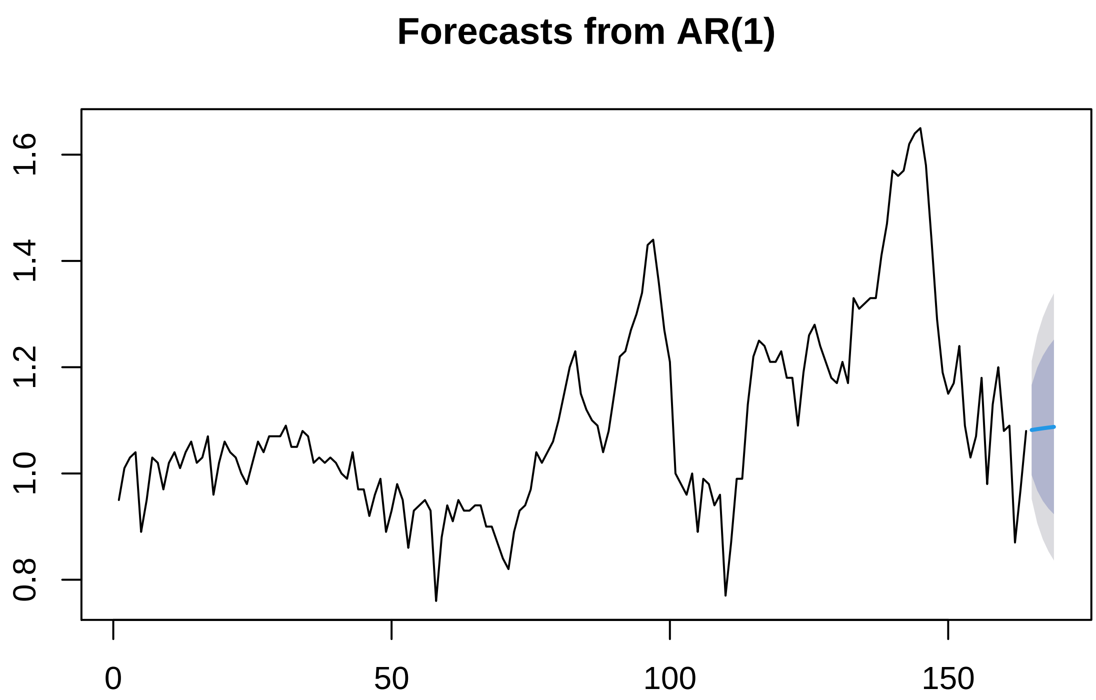
```r
print(avo.ar) # Here you can see the two coefficients of the two lags
```

```
##
## Call:
## ar(x = AveragePrice.ts[1:164], aic = FALSE, order.max = 1, method = "ols")
##
## Coefficients:
##      1
## 0.9248
##
## Intercept: 0.0008035 (0.005191)
##
## Order selected 1  sigma^2 estimated as  0.004393
```

```r
plot(AveragePrice.ts)
lines(time(AveragePrice.ts)[1:164],AveragePrice.ts[1:164] - avo.ar$resid, col = "green")
```



```r
plot(forecast(avo.ar, 5))
```




```r
# Forecast 5 steps-ahead
forecast(avo.ar, 5)
```

```
##     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 165       1.081781 0.9968424 1.166719 0.9518788 1.211682
## 166       1.083427 0.9677364 1.199118 0.9064934 1.260361
## 167       1.084950 0.9483452 1.221555 0.8760311 1.293869
## 168       1.086358 0.9341303 1.238586 0.8535457 1.319171
## 169       1.087660 0.9232451 1.252076 0.8362089 1.339112
```

```r
# Compare to the actual next 5
AveragePrice.ts[165:169]
```

```
## [1] 1.06 1.07 1.09 1.05 1.03
```

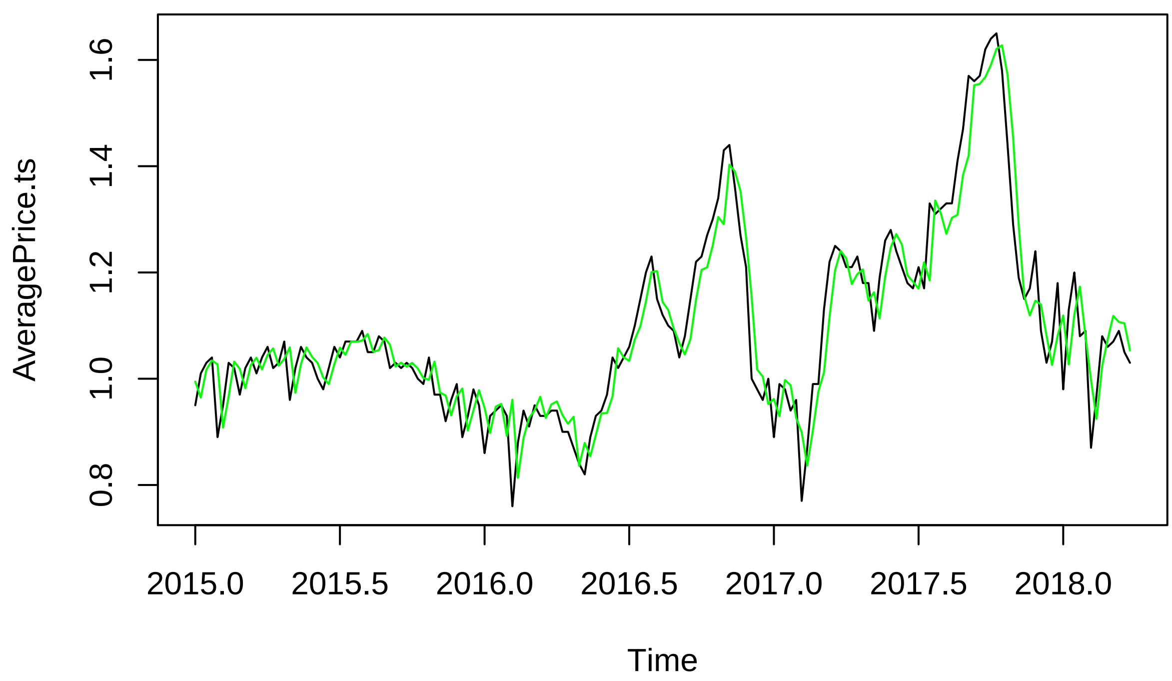
Using Seasonality With AR(1) Model

The ACF and PACF for average price show clear seasonality year-over-year, which is to be expected with produce such as avocados. As such, we investigated including seasonality in our model to account for this effect, with AR(1) and MA(1) seasonality.

```r
seasonal_model <- Arima(AveragePrice.ts, order = c(1,0,0), seasonal = list(order=c(1,0,1), period = 52))

summary(seasonal_model)
```

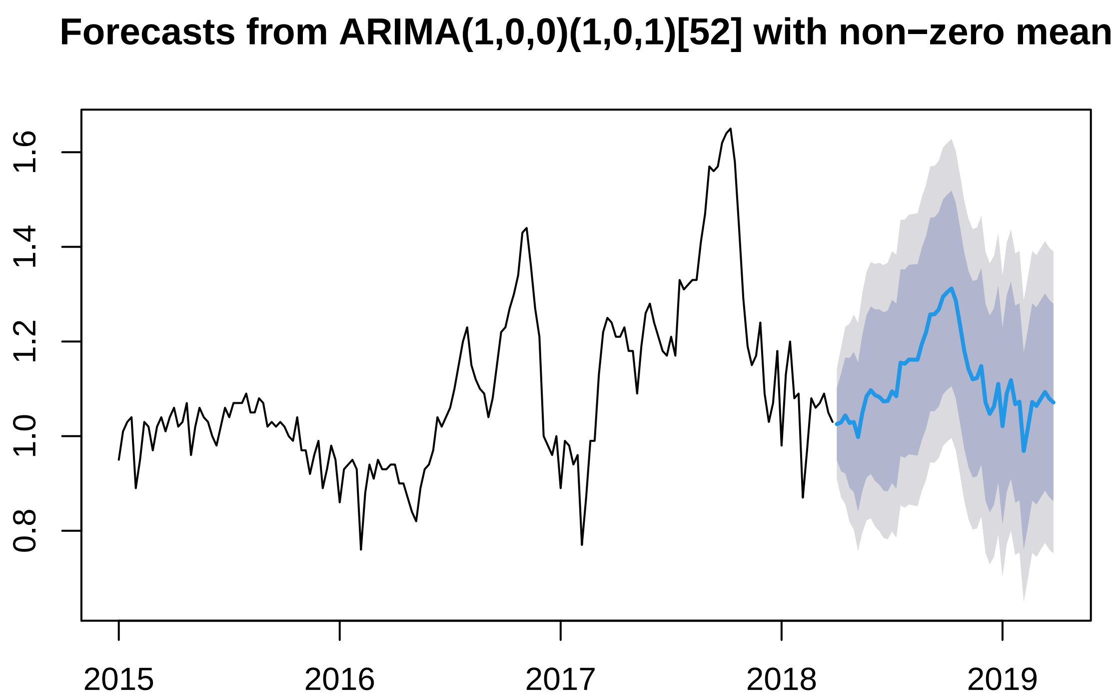
```
## Series: AveragePrice.ts
## ARIMA(1,0,0)(1,0,1)[52] with non-zero mean
##
## Coefficients:
##          ar1    sar1     sma1    mean
##       0.9318  0.5807  -0.1771  1.0888
## s.e.  0.0259  0.1897   0.2399  0.0944
##
## sigma^2 estimated as 0.003492:  log likelihood=233.37
## AIC=-456.73   AICc=-456.36   BIC=-441.08
##
## Training set error measures:
##                         ME       RMSE        MAE        MPE     MAPE      MASE
## Training set 0.001984865 0.05838771 0.04493045 -0.1104029 4.182295 0.2726588
##                    ACF1
## Training set 0.0883548
```

```r
plot(AveragePrice.ts)
lines(fitted(seasonal_model), col = "green")
```



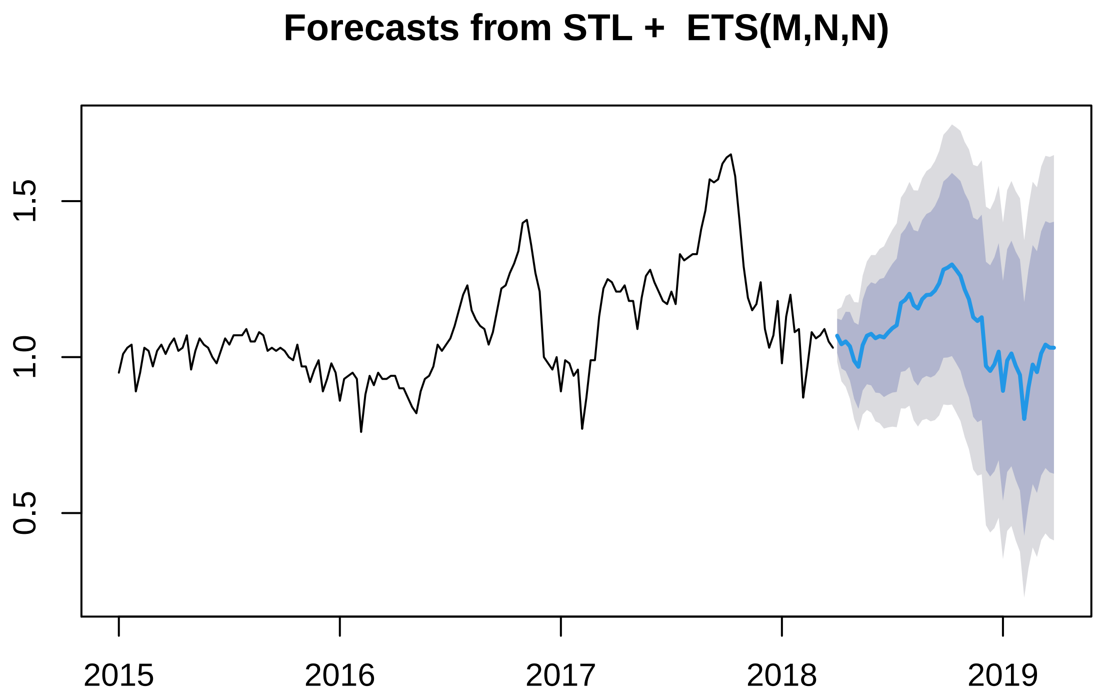
We can see that the fitted values follow the data very closely. Next, we use the seasonal model to predict average price for the next year.

```r
plot(forecast(seasonal_model,52))
```



We see that the predicted price follows the seasonal trend of rising thoughout the year, then falling at the end of the year as we expect. This matches up with our knowledge of avocados being cheapest in the winter months and rising in price during the rest of the year. We also looked into forecasting with the built-in exponential smoothing capabilities of the forecast() function.

```r
plot(forecast(AveragePrice.ts, 52))
```



The exponentially smoothed prediction also follows the seasonal trend, but predicts that the trend of average price will actually decrease in the next year, which we do not believe to be likely.

Buidling an ARDL model

Below is an ARDL(4,4) model. We built this model to see the statistical significance lags of both average price and total volume would have on predicting average price. We chose a lag of 4, for that represents one months worth of lags. The findings here are interesting in that lags of average price appeared to be less statistically significant than lags of volume. It appears that average price from 1 week ago, current total volume, and total volume from 1 week ago are statistically significant in explaining the current average price.

```r
# ARDL(4,4)
avo.ardl <- dynlm(AveragePrice.ts ~ L(AveragePrice.ts,1:4) + L(TotalVolume.ts, 0:4))
summary(avo.ardl)
```

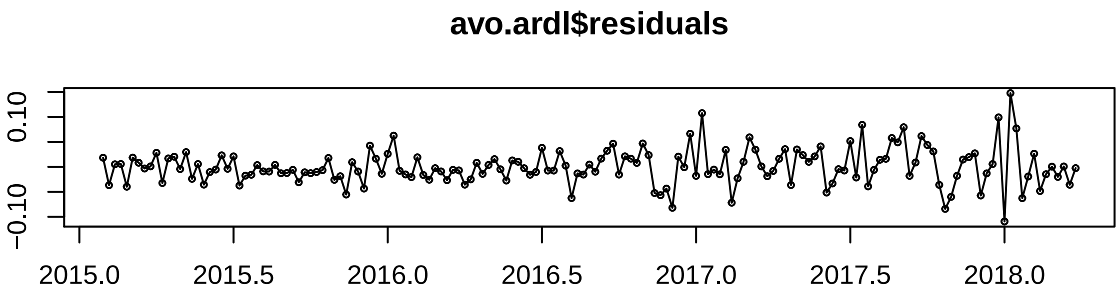
```
##
## Time series regression with "ts" data:
## Start = 2015(5), End = 2018(13)
##
## Call:
## dynlm(formula = AveragePrice.ts ~ L(AveragePrice.ts, 1:4) + L(TotalVolume.ts,
##     0:4))
##
## Residuals:
##       Min         1Q     Median         3Q        Max
```

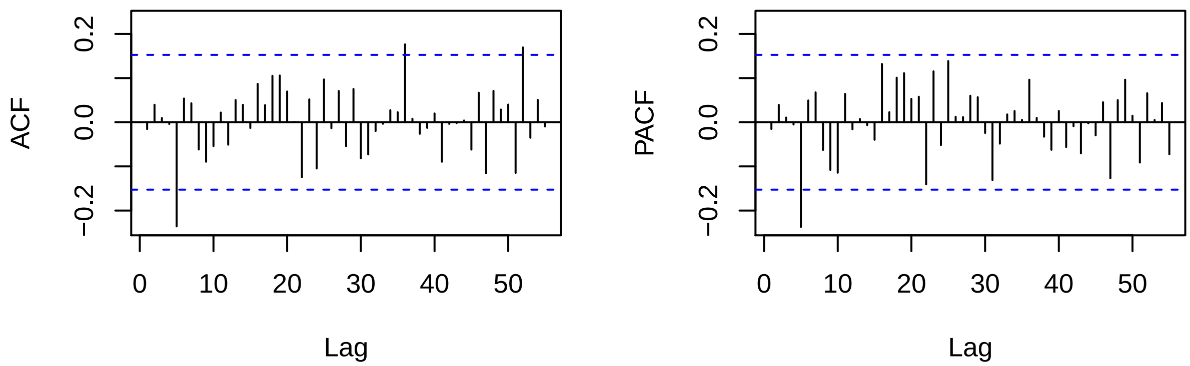
```
## -0.109239 -0.018026 -0.004676  0.018518  0.147490
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.115e-02  3.921e-02   0.794   0.4282
## L(AveragePrice.ts, 1:4)1  1.052e+00  7.891e-02  13.327  < 2e-16 ***
## L(AveragePrice.ts, 1:4)2  3.799e-02  1.146e-01   0.332   0.7406
## L(AveragePrice.ts, 1:4)3 -9.034e-02  1.135e-01  -0.796   0.4272
## L(AveragePrice.ts, 1:4)4 -5.069e-02  7.427e-02  -0.682   0.4960
## L(TotalVolume.ts, 0:4)0  -1.107e-08  6.383e-10 -17.346  < 2e-16 ***
## L(TotalVolume.ts, 0:4)1   8.821e-09  1.099e-09   8.023 2.38e-13 ***
## L(TotalVolume.ts, 0:4)2   2.898e-09  1.304e-09   2.222   0.0277 *
## L(TotalVolume.ts, 0:4)3  -9.564e-10  1.327e-09  -0.721   0.4721
## L(TotalVolume.ts, 0:4)4   1.086e-09  9.857e-10   1.102   0.2723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03678 on 155 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9561
## F-statistic: 398.3 on 9 and 155 DF,  p-value: < 2.2e-16
```

Testing for Serially Correlation

Here we test whether the ARDL model above violates the assumption that the errors are serially correlated. Looking at the ACF and PACF plots for the model's residuals, we do not see any distinctive pattern which suggests that the errors are not serially correlated. The Breusch-Godfrey test for higher order serial correlation also suggests that there is no serial correlation in the model, since the high p-value means that we fail the reject the null. This means that we do not have to correct for serial correlation.

```r
library(tseries)
tsdisplay(avo.ardl$residuals)
```






```r
bgtest(avo.ardl, order=1, type="F", fill=0)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  avo.ardl
## LM test = 1.0196, df1 = 1, df2 = 154, p-value = 0.3142
```

Testing Instrumental Variables

Using our subject mater expertise on avocados we thought to test total bags as an instrumental variable. The reason for this is we believed total bags to be a good indicator of total volume, but not necessarily of price. We created that IV test below. From our new model that takes into consideration total bags, it can be seen that there isn't much change in the significance of our parameters, suggesting that total bags is a weaker IV.

```r
# This will only run if the second
# Before testing total bags as an IV
no.iv.mod <- lm(AveragePrice ~ Total.Volume + Hass.Small + Hass.Large + Hass.Extra.Large, data = avocado
summary(no.iv.mod)
```

```
##
## Call:
## lm(formula = AveragePrice ~ Total.Volume + Hass.Small + Hass.Large +
```

```
##     Hass.Extra.Large, data = avocados_us_conventional)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25325 -0.07169 -0.00613  0.05348  0.35295
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.652e+00  5.303e-02  31.149  < 2e-16 ***
## Total.Volume      2.136e-09  3.070e-09   0.696    0.487
## Hass.Small       -8.550e-09  6.773e-09  -1.262    0.209
## Hass.Large       -3.567e-08  7.285e-09  -4.896 2.32e-06 ***
## Hass.Extra.Large -1.253e-07  3.075e-08  -4.074 7.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1161 on 164 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.5552
## F-statistic: 53.42 on 4 and 164 DF,  p-value: < 2.2e-16
```

```r
# Total Bags is the IV we are testing
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```r
total.bags.iv <- ivreg(AveragePrice ~ Total.Volume + Hass.Small + Hass.Large + Hass.Extra.Large | Hass.S
summary(total.bags.iv)
```

```
##
## Call:
## ivreg(formula = AveragePrice ~ Total.Volume + Hass.Small + Hass.Large +
##     Hass.Extra.Large | Hass.Small + Hass.Large + Hass.Extra.Large +
##     Total.Bags, data = avocados_us_conventional)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.253251 -0.071695 -0.006127  0.053476  0.352947
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.652e+00  5.303e-02  31.149  < 2e-16 ***
## Total.Volume      2.136e-09  3.070e-09   0.696    0.487
## Hass.Small       -8.550e-09  6.773e-09  -1.262    0.209
## Hass.Large       -3.567e-08  7.285e-09  -4.896 2.32e-06 ***
## Hass.Extra.Large -1.253e-07  3.075e-08  -4.074 7.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1161 on 164 degrees of freedom
## Multiple R-Squared: 0.5658,  Adjusted R-squared: 0.5552
## Wald test: 53.42 on 4 and 164 DF,  p-value: < 2.2e-16
```

Conclusions

Based on our analysis, we make the following key conclusions:

- Both average price and total volume follow an AR process, in which present values of the time series are related to past values.
- Current average price can be explained by recent lags in average price and total volume.
- We can predict the the short-term future average price given that it follows a seasonal pattern.
- The price of avocados throughout the year follows a consistent seasonal pattern of rising until a peak in early fall, then dropping to its lowest point at the end of the year. As such, we would advise consumers to avoid purchasing avocados in late summer and early fall, then get plenty of avocados during the winter for holiday and Super Bowl guacamole.
- This seasonal effect was weak in 2015 but became more prominent in following years. It's possible that external weather factors during the later years reduced the yield of out-of-season avocados, though the true reason is unclear in our data.

Future Work

For our future work, we believe we can improve on our model by testing the performance of our model. We can do this by using cross validation, in which we would divide our data into training and testing sets in order to evaluate how well our model performed. This evaluation can inform us if we should restructure our model to improve its accuracy.

We also would like to further investigate the difference between conventional and organic avocados. While conventional makes up the vast majority of the marketplace, we saw that the trends were different for the two types, so it would be interesting to look into the effect of the increasing popularity of organic produce on the avocado market.

In addition, we believe that it would useful to account for other variables not included in this dataset that could explain changes in avocado price and volume. For example, we could integrate information on weather patterns and economic conditions to provide context for our interpretations. This would help us explain the trends in the data within a bigger picture.

References

Avocado data from kaggle: https://www.kaggle.com/neuromusic/avocado-prices