**Avocado Prices – Mock analysis**

This analysis is a mock analysis of Avocado prices from 27.12.2015 up to 01.07.2018. During this analysis, I will ask questions that are geared towards learning and manipulating data. In doing so, the aim is two fold, both to learn, as well as keep record of progress being made. My interest in the subject, as well as the tools I’ve learned throughout the past two weeks, are mainly attributed to Sentdex. A data scientist with a YouTube channel that has left a positive impact on me.

The data available to us, other than the date is:

* Average price – this is a rolling average of the previous week
* Total volume – total amount of avocados sold
* 4046,4225,4770 – these are types of avocados , small, large and extra large
* Total bags – total amount of bags sold
* Type – conventional and organic
* Region – region where item was sold

## Initial observations and thoughts

* Observe the increase of average price over the years
* Observe the seasonal demand for avocados
* Correlation between the amount sold and the average price of the avocados
* Quick online search to see any important news between the dates that might have had a direct effect on avocado prices
* What is the price difference between conventional and organic avocados
* Compare prices between regions

We will look into this information, and take time to explore any new observations while answering the questions above.

## Initial routine and data preparing

Importing the libraries that are most commonly used for analysis. Being a basic analysis, we will be importing numpy, pandas, seaborn and matplotlib, followed by importing the dataset. The data frame is typically named ‘df’

import numpy as np

import pandas as pd

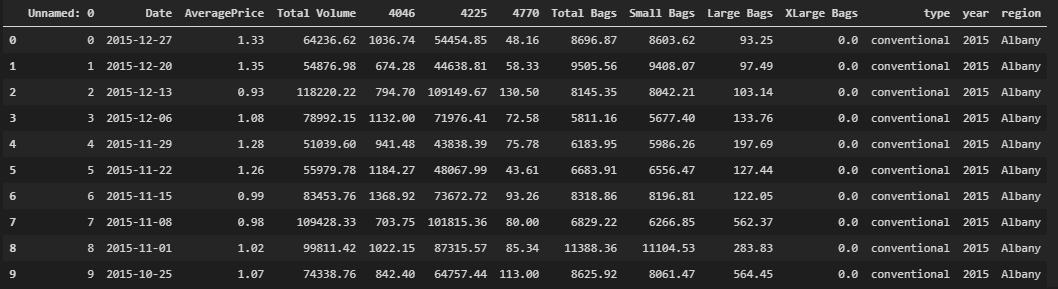
import seaborn as sb

import matplotlib.pyplot as plt

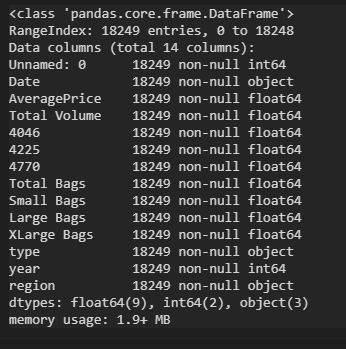
df = pd.read\_csv('avocado.csv')

The first thing we want is to have a quick look at the first values

df.head(10)



* The index is purely numerical, we can set this to date as most of our data will be plotted against date
* We only have the region ‘Albany’ here, we will need to look into the general order of data.
* We will need to check that the data is in the correct format.
  + df.info()

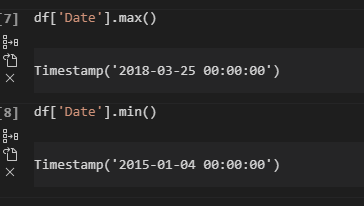


* Date will need to be converted to datetime.

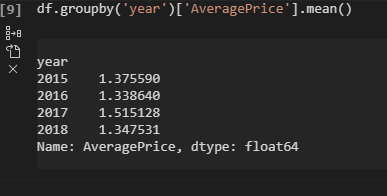
df['Date'] =  pd.to\_datetime(df['Date'])



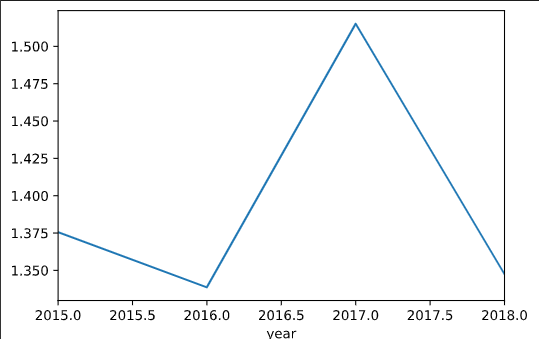
* Looking at the first date and last date
* df['Date'].max()
* df['Date'].min()



* Let’s check the average of all avocados per year
* df.groupby('year')['AveragePrice'].mean()



average\_year\_price.plot(figsize=(15,11))



## California Vs Los Angelis

The first analysis I will carry out is comparing the prices of conventional avocados and organic avocados. At face value, organic prices should typically produce a higher average price. Naturally this would be ignoring many other factors such as supply and demand. Avocados are generally looked at as a ‘luxury good’ and as such will be more sensitive to external factors.

The dataset has various regions, to keep things simple; we will look at two states, California and Los Angelis. California was chosen as it is one of the country’s largest avocado growers, and Los Angelis was chosen at random. We will also only look at 2015 for the time being to limit the dataset.

The first step is to list the conditions. These can be hardcoded into the code, however both in terms of practice as well as to create something that is reusable and reproducible, we will create conditions as variables.

con\_conventional = df['type'] == 'conventional'

con\_organic = df['type']=='organic'

con\_2015 = df['year'] == 2015

con\_region\_calif = df['region']== 'California'

con\_region\_LosAngeles = df['region'] == 'LosAngeles'

Creating the dataframe for Average Price, in 2015, in California and are conventional & organic.

average\_15\_conv\_calif = df[con\_conventional & con\_2015 & con\_region\_calif]

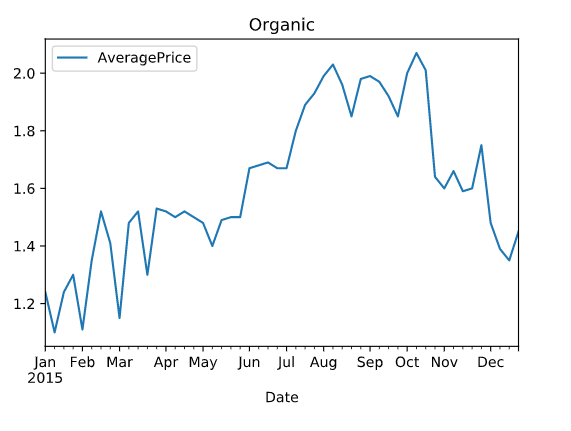
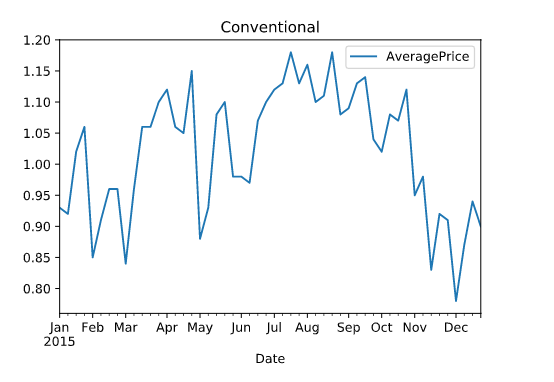
average\_15\_conv\_calif.sort\_values('Date')

average\_15\_conv\_calif.plot(y='AveragePrice',x='Date')

average\_15\_orga\_calif = df[con\_organic & con\_2015 & con\_region\_calif]

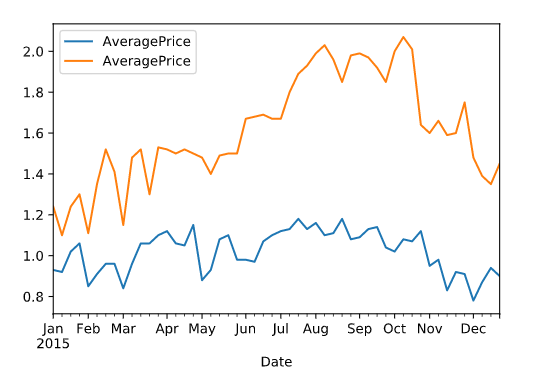
average\_15\_orga\_calif.sort\_values('Date')

average\_15\_orga\_calif.plot(y='AveragePrice',x='Date')

 Joining both graphs together

ax=average\_15\_conv\_calif.plot(y='AveragePrice',x='Date')

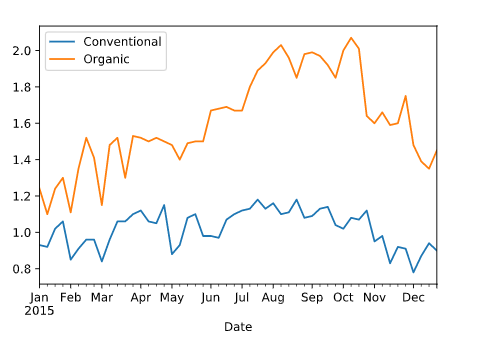
average\_15\_orga\_calif.plot(ax=ax,y='AveragePrice',x='Date')



Great, let’s give it some labels

ax=average\_15\_conv\_calif.plot(y='AveragePrice',x='Date',label='Conventional')

average\_15\_orga\_calif.plot(ax=ax,y='AveragePrice',x='Date',label='Organic')



Ok, we have some data visualised here and we can carry out some very simple (and somewhat inaccurate in the grander scheme of things) analysis. Throughout the year it is clear that the prices of organic avocados is higher than the prices of conventional ones, this was expected and we can now prove it too.

Throughout the year the movement in price has been similar between organic and conventional, with organic prices movement slightly trailing behind the organic on the most part. An increase in price from June for both avocado types could be as a result of shortage of avocadoes which was widely documented in May. It would be interesting to check these figures against another year to see if it is a seasonal fluctuation or an event driven change. Let’s look at 2016.

average\_16\_conv\_calif= df[con\_conventional & con\_2016 & con\_region\_calif]

average\_16\_orga\_calif = df[con\_organic & con\_2016 & con\_region\_calif]

average\_16\_conv\_calif.sort\_values('Date')

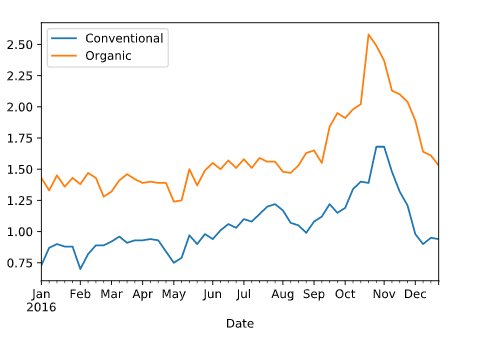
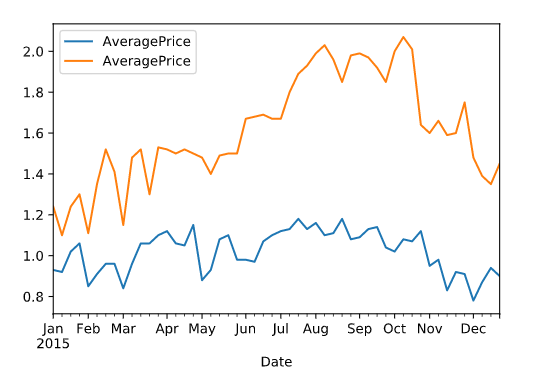
average\_16\_orga\_calif.sort\_values('Date')

average\_16\_conv\_calif.plot(y='AveragePrice',x='Date', title='Conventional')

average\_16\_orga\_calif.plot(y='AveragePrice',x='Date',title='Organic')

ax=average\_16\_conv\_calif.plot(y='AveragePrice',x='Date',label='Conventional')

average\_16\_orga\_calif.plot(ax=ax,y='AveragePrice',x='Date',label='Organic')



Above we have similar figures for 2016. An important thing to note is that with the conditions set up as variables, it was much quicker to extract the data. The analysis – There is a gradual price increase however less so than 2015. Prices definitely increase towards the end of the year – this could be attributable to the fact that avocados enjoy warmer weather to grow, lower temperatures may create a scarcity.

Bringing in Los Angelis, we will transpose all the above for **Los Angelis**

# Los Angelis for 2015

average\_15\_conv\_la = df[con\_conventional & con\_2015 & con\_region\_la]

average\_15\_conv\_la.sort\_values('Date')

average\_15\_conv\_la.plot(y='AveragePrice',x='Date', title='Conventional')

average\_15\_orga\_la = df[con\_organic & con\_2015 & con\_region\_la]

average\_15\_orga\_la.sort\_values('Date')

average\_15\_orga\_la.plot(y='AveragePrice',x='Date',title='Organic')

ax=average\_15\_conv\_la.plot(y='AveragePrice',x='Date',label='Conventional')

average\_15\_orga\_la.plot(ax=ax,y='AveragePrice',x='Date',label='Organic')

# Los Angelis for 2016

average\_16\_conv\_la= df[con\_conventional & con\_2016 & con\_region\_la]

average\_16\_orga\_la = df[con\_organic & con\_2016 & con\_region\_calif]

average\_16\_conv\_la.sort\_values('Date')

average\_16\_orga\_la.sort\_values('Date')

average\_16\_conv\_la.plot(y='AveragePrice',x='Date', title='Conventional')

average\_16\_orga\_la.plot(y='AveragePrice',x='Date',title='Organic')

ax=average\_16\_conv\_la.plot(y='AveragePrice',x='Date',label='Conventional')

average\_16\_orga\_la.plot(ax=ax,y='AveragePrice',x='Date',label='Organic')

