Avocado Price Analysis

February 1, 2021

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy import stats
```

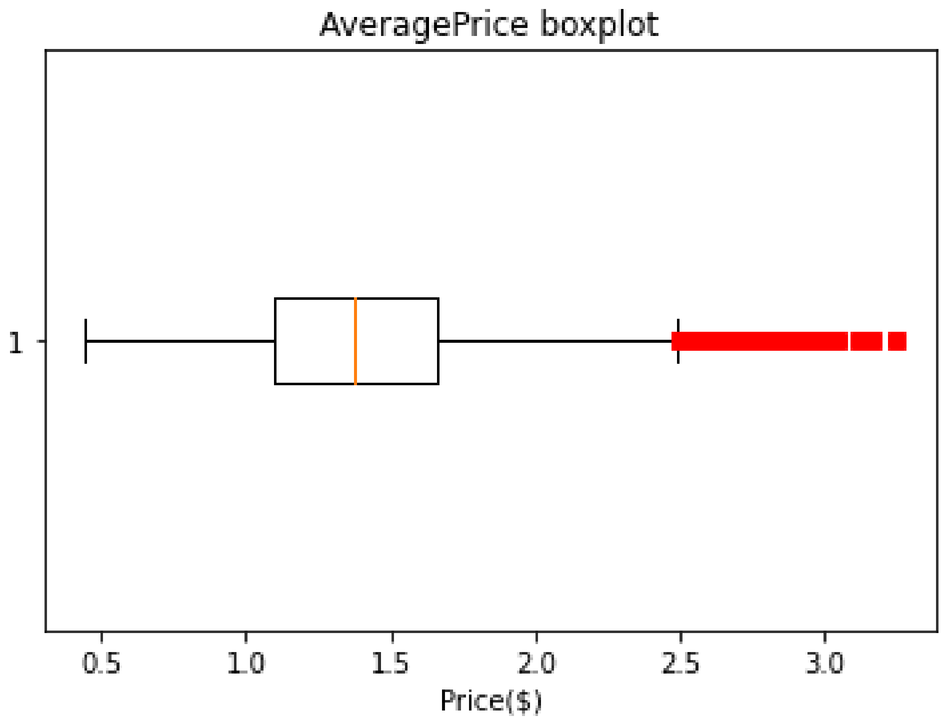
[2]:
```
data = pd.read_csv('avocado.csv',delimiter = ',')
data.drop(columns='Unnamed: 0',inplace=True)
```

[4]:
```
data.info()
```

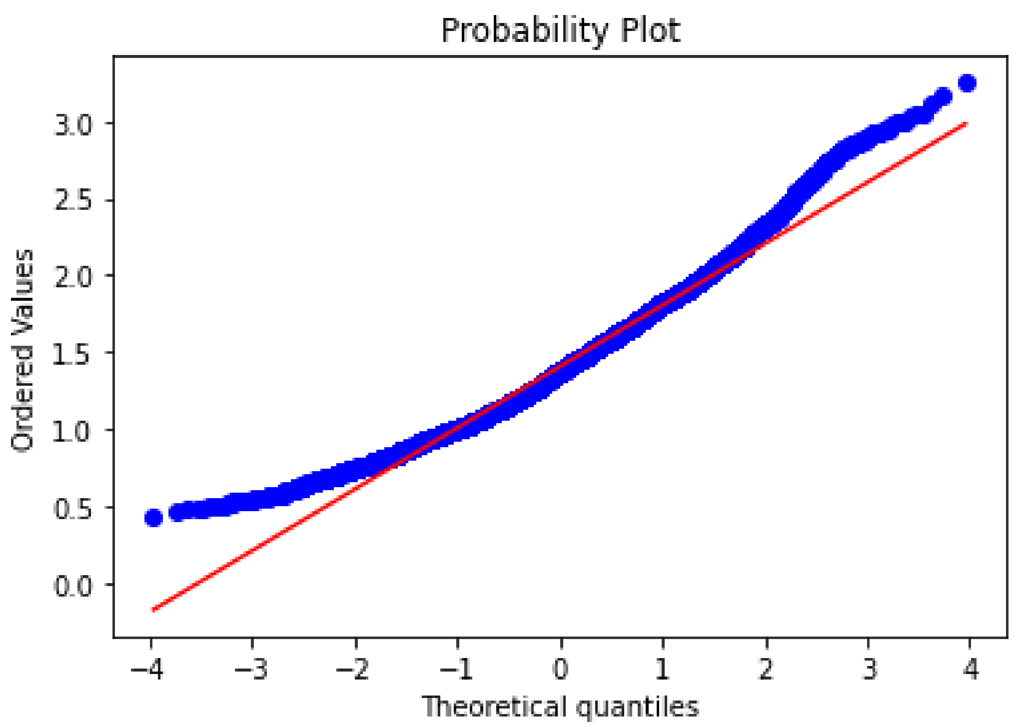
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18249 entries, 0 to 18248
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Date          18249 non-null  object
 1   AveragePrice  18249 non-null  float64
 2   Total Volume  18249 non-null  float64
 3   4046          18249 non-null  float64
 4   4225          18249 non-null  float64
 5   4770          18249 non-null  float64
 6   Total Bags    18249 non-null  float64
 7   Small Bags    18249 non-null  float64
 8   Large Bags    18249 non-null  float64
 9   XLarge Bags   18249 non-null  float64
 10  type          18249 non-null  object
 11  year          18249 non-null  int64
 12  region        18249 non-null  object
dtypes: float64(9), int64(1), object(3)
memory usage: 1.8+ MB
```

[120]:
```
plt.boxplot(data['AveragePrice'], 0, 'rs', 0)
plt.title('AveragePrice boxplot')
plt.xlabel('Price($)')
plt.plot()
```

[120]: []



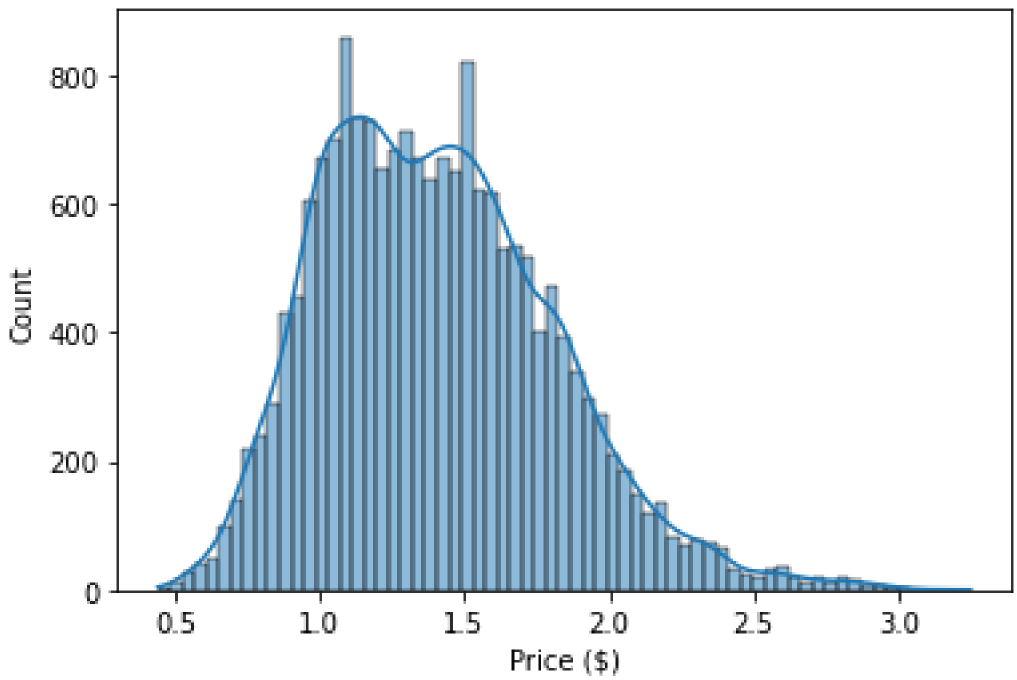
[6]:
```
ax4 = plt.subplot()
res = stats.probplot(data['AveragePrice'],dist ='norm',plot=plt)
```



Not a great fit for a normal distriibution.

[132]:
```
sns.histplot(data['AveragePrice'],kde=True)
plt.xlabel('Price ($)')
plt.plot()
```

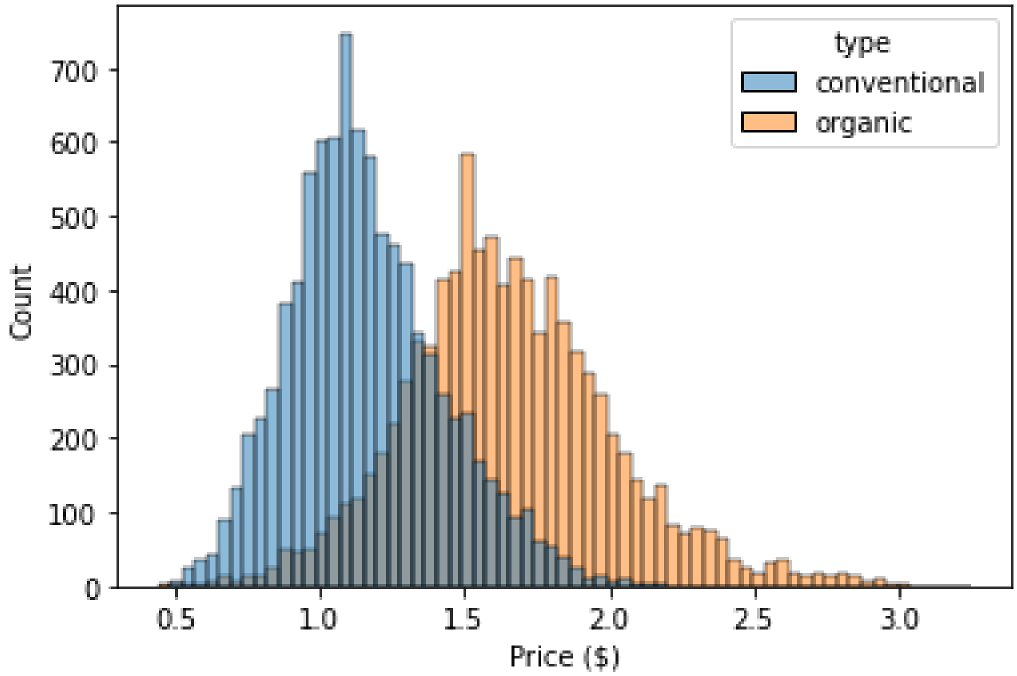
[132]: []



It appears to be a bimodal distribution, which is strange for a price of the same item you would expect it to be unimodal.

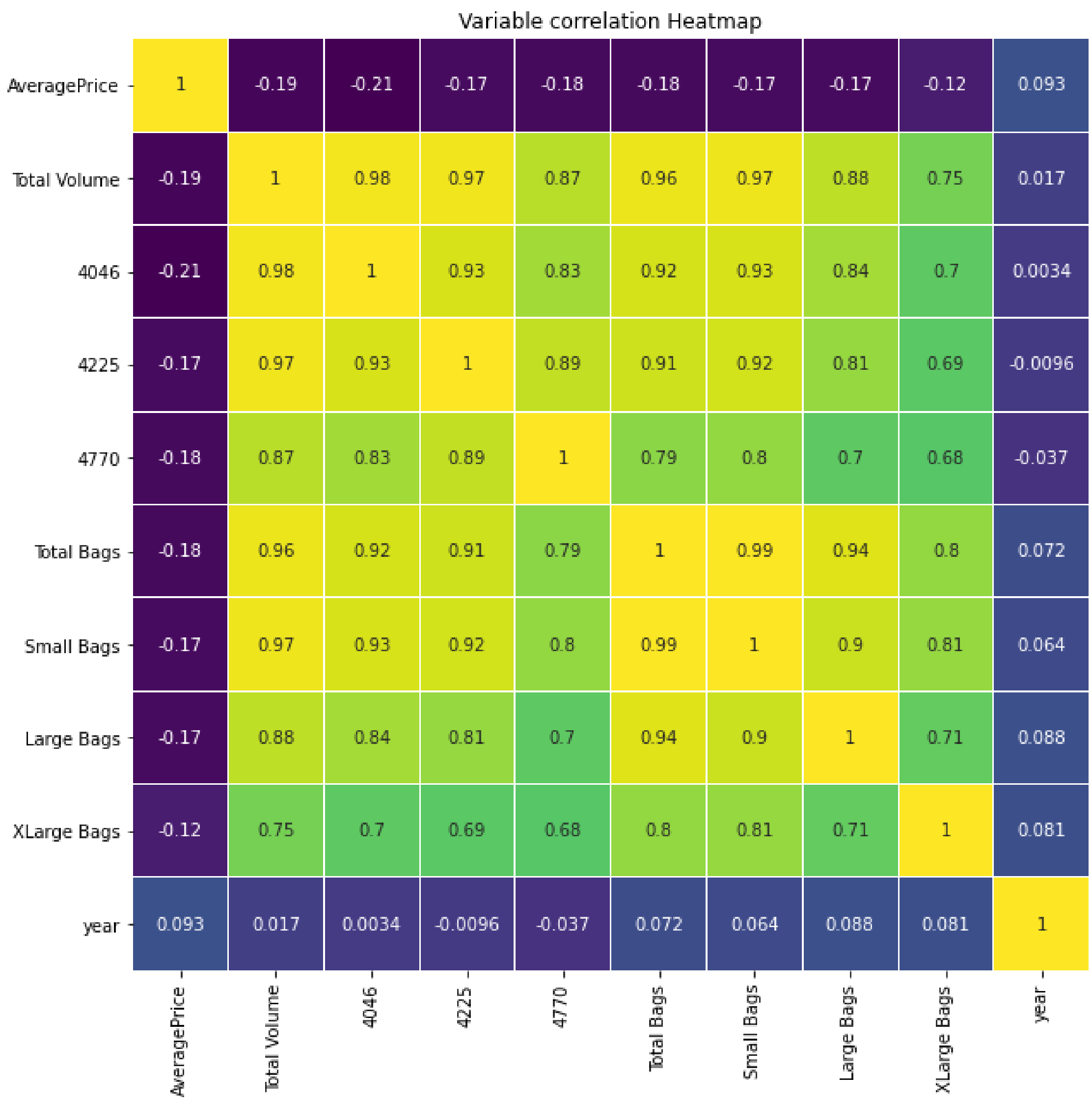
later in the analysis, I come to compare the variable 'type' which categorizes the avocados between organic and conventional. this explains the bimodal distribution, as the skews for organic and conv are different

[131]:
```
sns.histplot(data,x='AveragePrice',hue='type')
plt.xlabel('Price ($)')
plt.show()
```



[126]:

```
d_corr = data.corr()
fig,ax = plt.subplots(figsize=(10,10))
ax = sns.heatmap(d_corr,annot=True,linewidth=0.01,cbar=False,cmap='viridis')
ax.set_title('Variable correlation Heatmap')
plt.show()
```

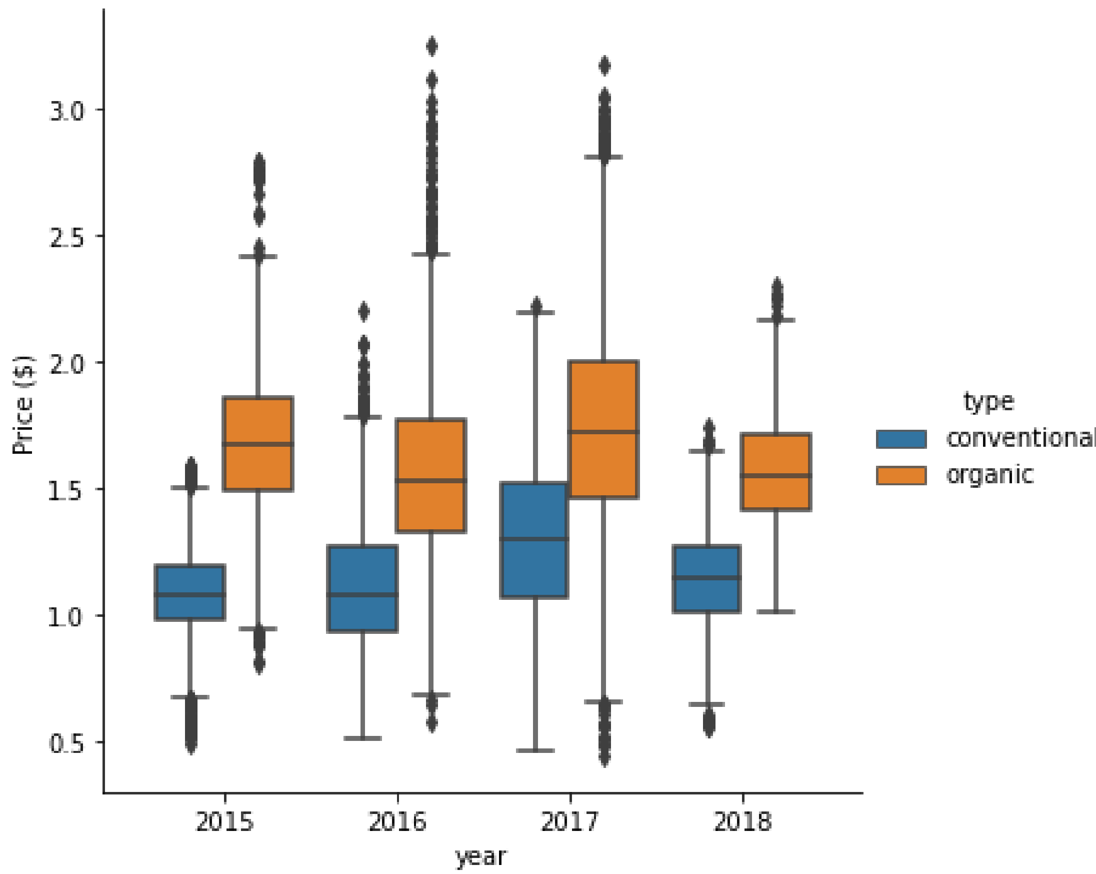


0.1 Categorical

0.1.1 Type organic / conventional

[141]:
```
sns.catplot(x ='year',y='AveragePrice', data=data,hue='type',kind='box')
plt.ylabel('Price ($)')
plt.plot()
```

[141]: []

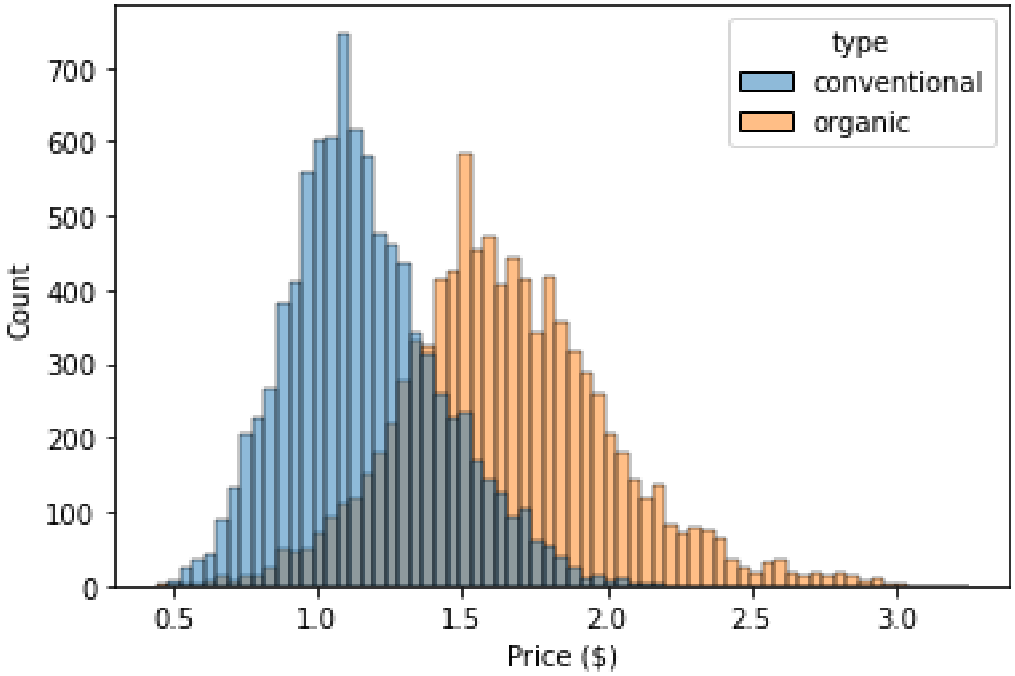


- Organic avocados tend to be more expensive that conventional avocados.

bernulli 1-organic, 0-conventional

[143]:
```
sns.histplot(data,x='AveragePrice',hue='type')
plt.xlabel('Price ($)')
plt.plot()
```

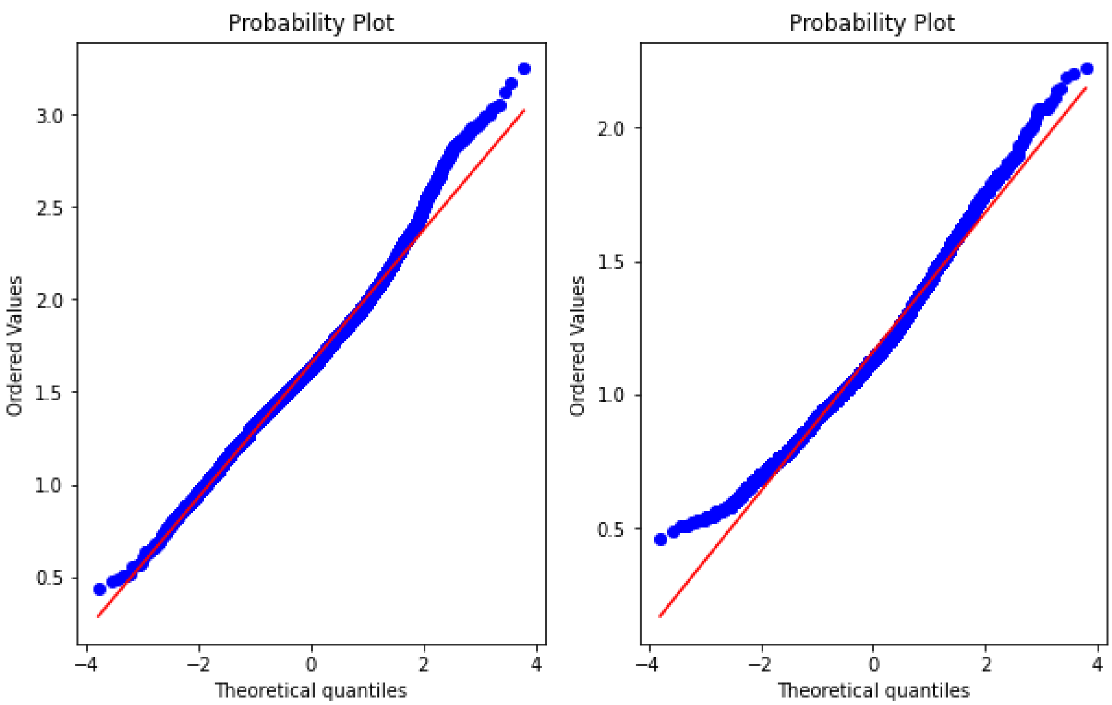
[143]: []



distribution fit

[87]:
```
f, ax = plt.subplots(1,2,figsize=(10,6))

stats.probplot(data.loc[data['type']=='organic','AveragePrice'],dist
 ↪='norm',plot=ax[0])
stats.probplot(data.loc[data['type']=='conventional','AveragePrice'],dist
 ↪='norm',plot=ax[1])
plt.plot()
```

[87]: []



the spread of values for organic is much larger. 4Q >3.0 compared to 4Q<3.0

[102]:
```
avo_type  = data.groupby('type').agg(['mean','std','count'])
avo_type = round(avo_type,2)
#mean diff
mean_diff =␣
 ↪avo_type['AveragePrice']['mean']['organic']-avo_type['AveragePrice']['mean']['conventional']
#standard deviation
organic_s1 = avo_type['AveragePrice']['std']['organic']
conventional_s2 = avo_type['AveragePrice']['std']['conventional']
print('mean difference',mean_diff,'p')
```

```
mean difference 0.49 p
```

X_1 = Organic Avocados, AveragePrice dist.

X_2 = Conventional Avocados, AveragePrice dist.

$\bar{X_1} - \bar{X_2} = 0.49$

$\alpha = 0.01$

$$\sigma^2_{\bar{X_1}+\bar{X_2}} \approx \frac{S_1}{n_1} + \frac{S_2}{n_2}$$

[117]:
```
sampling_stderr = (organic_s1/
 ↪avo_type['AveragePrice']['count']['organic'])+(conventional_s2/
 ↪avo_type['AveragePrice']['count']['conventional'])
sampling_stderr = np.sqrt(sampling_stderr)
print('sampling std error approx',sampling_stderr)

crit_limit =2.33
limit = crit_limit*sampling_stderr
print(round(mean_diff - limit,2),'to',round(mean_diff +limit,2))
```

```
sampling std error approx 0.008243223411136655
0.47 to 0.51
```

Confident (That the true Mean difference in price between Organic avocados and conventional avocados is between 0.47p and 0.51p) $\approx 99\%$

0.2 Region

Groupby region, with lambda agg to reduce oulier prices.

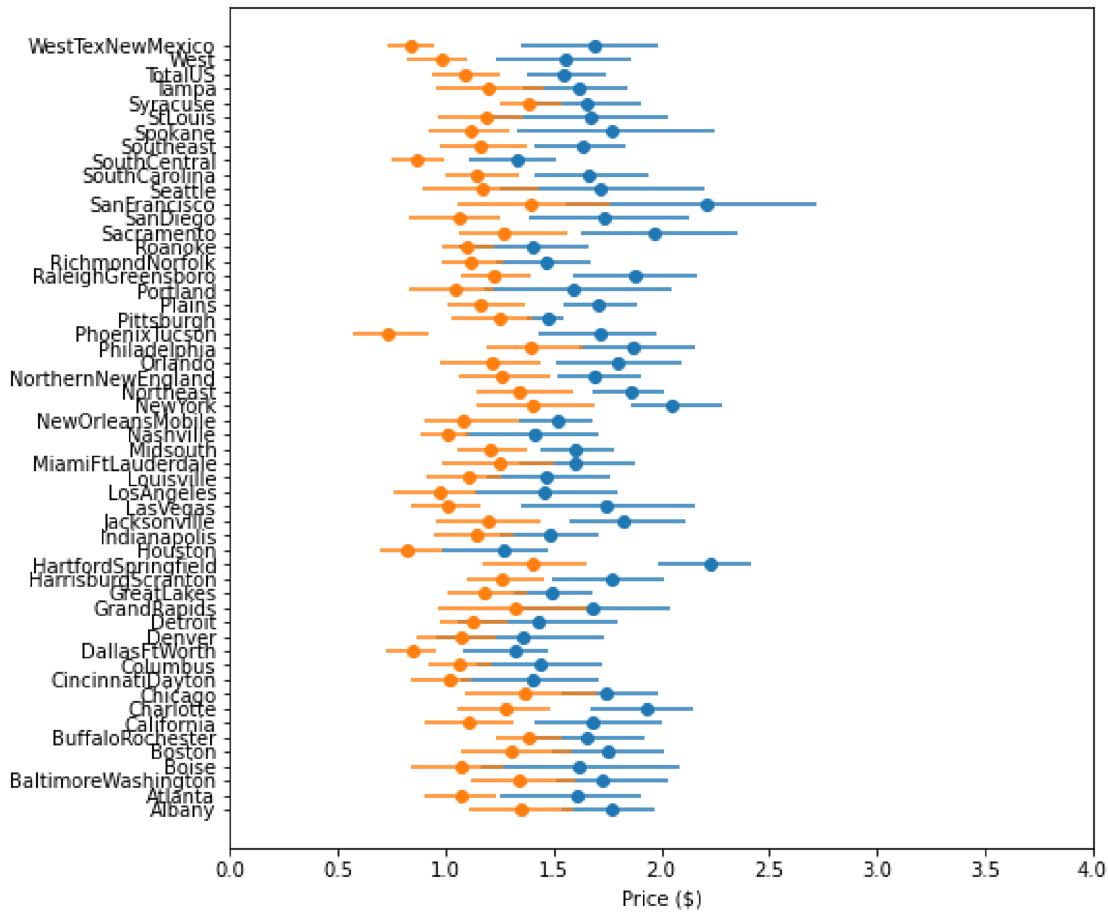
[351]:
```
region_data = data.groupby(['region','type'])['AveragePrice'].agg(['mean',
                                                            lambda x : np.
 ↪quantile(x,.15),
                                                            lambda x : np.
 ↪quantile(x,.85)])
region_data = region_data.unstack()
region_data.rename(columns={'<lambda_0>':'q15','<lambda_1>':'q85'},inplace=True)
err_pdata = region_data.copy()
region_data.head()
```

[351]:
```
                            mean                    q15                 q85  \
type                conventional   organic conventional organic conventional
region
Albany                  1.348757  1.773314        1.110    1.54        1.588
Atlanta                 1.068817  1.607101        0.902    1.25        1.230
BaltimoreWashington     1.344201  1.724260        1.120    1.51        1.600
Boise                   1.076036  1.620237        0.836    1.16        1.268
Boston                  1.304379  1.757396        1.070    1.49        1.580


type                organic
region
Albany                1.970
Atlanta               1.906
BaltimoreWashington   2.028
Boise                 2.088
Boston                2.010
```

[352]:
```
#reformating the quartile columns for the error bar plot
err_pdata['q15'] = err_pdata['mean'] - err_pdata['q15']
err_pdata['q85']= err_pdata['q85'] - err_pdata['mean']
err_pdata.head()
```

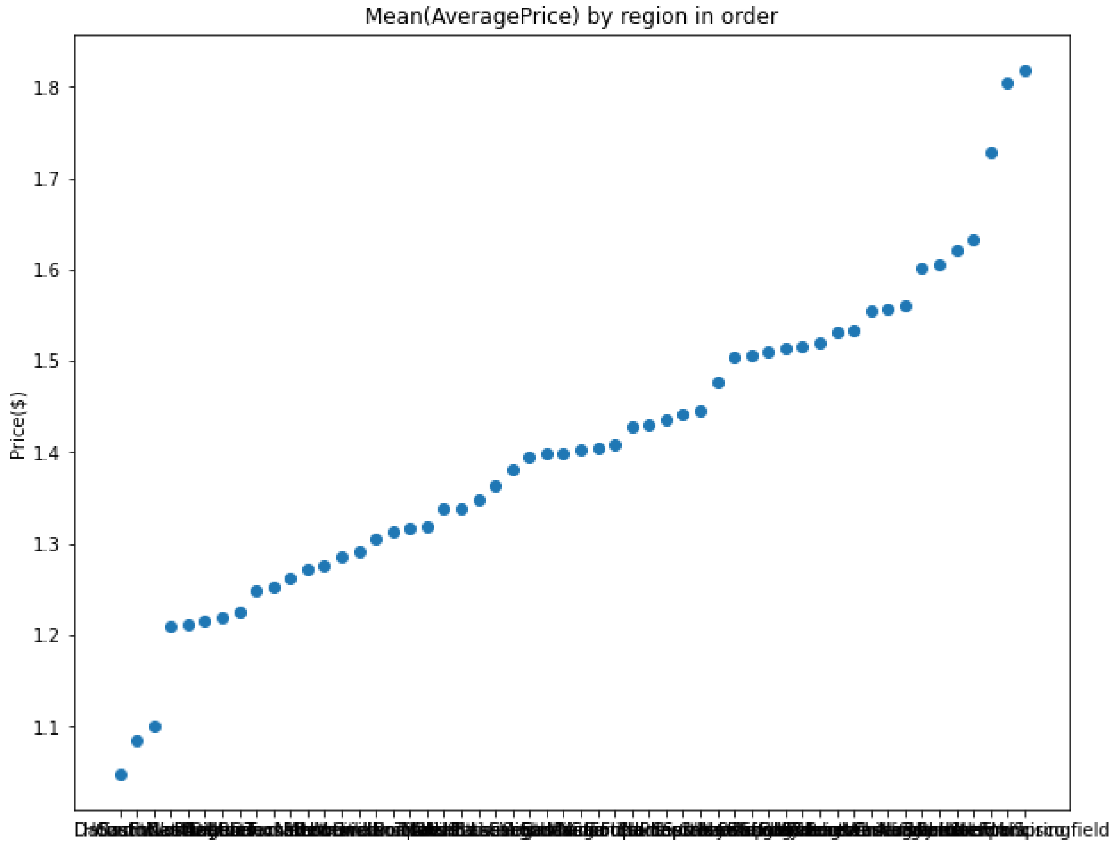
[352]:
```
                             mean                      q15             \
type                 conventional   organic conventional   organic
region
Albany                   1.348757  1.773314     0.238757  0.233314
Atlanta                  1.068817  1.607101     0.166817  0.357101
BaltimoreWashington      1.344201  1.724260     0.224201  0.214260
Boise                    1.076036  1.620237     0.240036  0.460237
Boston                   1.304379  1.757396     0.234379  0.267396

                              q85
type                 conventional   organic
region
Albany                   0.239243  0.196686
Atlanta                  0.161183  0.298899
BaltimoreWashington      0.255799  0.303740
Boise                    0.191964  0.467763
Boston                   0.275621  0.252604
```

[279]:
```
fig,ax = plt.subplots(figsize=(8,8))
ax.errorbar(y = test.
 ↪index,x=test['mean']['organic'],xerr=[err_pdata['q15']['organic'],err_pdata['q85']['organic
 ↪='o')
ax.errorbar(y = test.
 ↪index,x=test['mean']['conventional'],xerr=[err_pdata['q15']['conventional'],err_pdata['q85']
 ↪='o')
ax.set_xticks([0,0.5,1,1.5,2,2.5,3,3.5,4])
ax.set_xlabel('Price ($)')
plt.show()
```



[355]:
```python
price_by_region = data.groupby('region')['AveragePrice'].agg(['mean'])
price_by_region = price_by_region.sort_values('mean')
price_by_region = price_by_region.reset_index()
fig,ax = plt.subplots(figsize = (10,8))
ax.scatter(x=price_by_region['region'],y=price_by_region['mean'])
ax.set_title('Mean(AveragePrice) by region in order')
ax.set_ylabel('Price($)')
plt.plot()
```

[355]: []


Mean(AveragePrice) by region in order
Price($)
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
Springfield

[ ]: