# EXPLORATORY DATA ANALYSIS IN AVOCADO PRICES

In [1]:
```
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
```

In [7]:
```
import warnings
warnings.filterwarnings("ignore")
```

## Creating DataFrames

In [2]:
```
list_of_dicts = [
     {"name": "Ginger", "breed": "Dachshund", "height_cm": 22,"weight_kg": 10, "dat
    {"name": "Scout", "breed": "Dalmatian", "height_cm": 59,"weight_kg": 25, "date_
]
new_dogs = pd.DataFrame(list_of_dicts)
new_dogs
```

Out[2]:

| | name | breed | height_cm | weight_kg | date_of_birth |
|---|---|---|---|---|---|
| **0** | Ginger | Dachshund | 22 | 10 | 2019-03-14 |
| **1** | Scout | Dalmatian | 59 | 25 | 2019-05-09 |

In [3]:
```
dict_of_lists = {
     "name": ["Ginger", "Scout"],
     "breed": ["Dachshund", "Dalmatian"],
     "height_cm": [22, 59],
     "weight_kg": [10, 25],
     "date_of_birth": ["2019-03-14","2019-05-09"]  }
new_dogs = pd.DataFrame(dict_of_lists)
new_dogs
```

Out[3]:

| | name | breed | height_cm | weight_kg | date_of_birth |
|---|---|---|---|---|---|
| **0** | Ginger | Dachshund | 22 | 10 | 2019-03-14 |
| **1** | Scout | Dalmatian | 59 | 25 | 2019-05-09 |

In [4]:
```
avocado = pd.read_csv(r"C:\Users\JANHAVI\Desktop\avocado.csv")
```

In [5]:
```
avocado.head()
```

Out[5]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 | 93.25 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 | 97.49 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 | 103.14 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 | 133.76 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 | 197.69 |

In [8]:
```
avocado = pd.read_csv(r"C:\Users\JANHAVI\Desktop\avocado.csv",parse_dates=True, inc
avocado.head()
```

Out[8]:

| Date | Unnamed: 0 | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags |
|---|---|---|---|---|---|---|---|---|---|
| **2015-12-27** | 0 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 | 93.25 |
| **2015-12-20** | 1 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 | 97.49 |
| **2015-12-13** | 2 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 | 103.14 |
| **2015-12-06** | 3 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 | 133.76 |
| **2015-11-29** | 4 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 | 197.69 |

In [9]:
```
avocado = avocado.reset_index(drop=True)
avocado.head()
```

Out[9]:

| | Unnamed: 0 | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags | XLarg Bag |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 | 93.25 | 0 |
| **1** | 1 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 | 97.49 | 0 |
| **2** | 2 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 | 103.14 | 0 |
| **3** | 3 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 | 133.76 | 0 |
| **4** | 4 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 | 197.69 | 0 |

In [10]:
```
avocado.to_csv("test_write.csv")
```

In [12]:
```
avocado = pd.read_csv(r"C:\Users\JANHAVI\Desktop\avocado.csv")
avocado.head()
```

Out[12]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 | 93.25 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 | 97.49 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 | 103.14 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 | 133.76 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 | 197.69 |

In [13]:
```
avocado.tail(10)
```

Out[13]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | La B |
|---|---|---|---|---|---|---|---|---|---|---|
| **18239** | 2 | 11-03-2018 | 1.56 | 22128.42 | 2162.67 | 3194.25 | 8.93 | 16762.57 | 16510.32 | 252 |
| **18240** | 3 | 04-03-2018 | 1.54 | 17393.30 | 1832.24 | 1905.57 | 0.00 | 13655.49 | 13401.93 | 253 |
| **18241** | 4 | 25-02-2018 | 1.57 | 18421.24 | 1974.26 | 2482.65 | 0.00 | 13964.33 | 13698.27 | 266 |
| **18242** | 5 | 18-02-2018 | 1.56 | 17597.12 | 1892.05 | 1928.36 | 0.00 | 13776.71 | 13553.53 | 223 |
| **18243** | 6 | 11-02-2018 | 1.57 | 15986.17 | 1924.28 | 1368.32 | 0.00 | 12693.57 | 12437.35 | 256 |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 | 431 |
| **18245** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.84 | 8940.04 | 324 |
| **18246** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.11 | 9351.80 | 42 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 | 50 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 | 26 |

In [14]:
```
avocado.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18249 entries, 0 to 18248
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    18249 non-null  int64
 1   Date          18249 non-null  object
 2   AveragePrice  18249 non-null  float64
 3   Total Volume  18249 non-null  float64
 4   4046          18249 non-null  float64
 5   4225          18249 non-null  float64
 6   4770          18249 non-null  float64
 7   Total Bags    18249 non-null  float64
 8   Small Bags    18249 non-null  float64
 9   Large Bags    18249 non-null  float64
 10  XLarge Bags   18249 non-null  float64
 11  type          18249 non-null  object
 12  year          18249 non-null  int64
 13  region        18249 non-null  object
dtypes: float64(9), int64(2), object(3)
memory usage: 1.9+ MB
```

In [15]:
```
print(avocado.shape)
```

```
(18249, 14)
```

In [16]:
```
avocado.describe()
```

Out[16]:

| | Unnamed: 0 | AveragePrice | Total Volume | 4046 | 4225 | 4770 | To |
|---|---|---|---|---|---|---|---|
| **count** | 18249.000000 | 18249.000000 | 1.824900e+04 | 1.824900e+04 | 1.824900e+04 | 1.824900e+04 | 1.824 |
| **mean** | 24.232232 | 1.405978 | 8.506440e+05 | 2.930084e+05 | 2.951546e+05 | 2.283974e+04 | 2.396 |
| **std** | 15.481045 | 0.402677 | 3.453545e+06 | 1.264989e+06 | 1.204120e+06 | 1.074641e+05 | 9.862 |
| **min** | 0.000000 | 0.440000 | 8.456000e+01 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000 |
| **25%** | 10.000000 | 1.100000 | 1.083858e+04 | 8.540700e+02 | 3.008780e+03 | 0.000000e+00 | 5.088 |
| **50%** | 24.000000 | 1.370000 | 1.073768e+05 | 8.645300e+03 | 2.906102e+04 | 1.849900e+02 | 3.974 |
| **75%** | 38.000000 | 1.660000 | 4.329623e+05 | 1.110202e+05 | 1.502069e+05 | 6.243420e+03 | 1.107 |
| **max** | 52.000000 | 3.250000 | 6.250565e+07 | 2.274362e+07 | 2.047057e+07 | 2.546439e+06 | 1.937 |

In [17]:
```
avocado.values
```

Out[17]:
```
array([[0, '27-12-2015', 1.33, ..., 'conventional', 2015, 'Albany'],
       [1, '20-12-2015', 1.35, ..., 'conventional', 2015, 'Albany'],
       [2, '13-12-2015', 0.93, ..., 'conventional', 2015, 'Albany'],
       ...,
       [9, '21-01-2018', 1.87, ..., 'organic', 2018, 'WestTexNewMexico'],
       [10, '14-01-2018', 1.93, ..., 'organic', 2018, 'WestTexNewMexico'],
       [11, '07-01-2018', 1.62, ..., 'organic', 2018, 'WestTexNewMexico']],
      dtype=object)
```

In [18]:
```
print(avocado.columns)
```

```
Index(['Unnamed: 0', 'Date', 'AveragePrice', 'Total Volume', '4046', '4225',
       '4770', 'Total Bags', 'Small Bags', 'Large Bags', 'XLarge Bags', 'type',
       'year', 'region'],
      dtype='object')
```

# Appending & Concatenating Series

In [21]:
```
import pandas as pd

even = pd.Series([2,4,6,8,10])
odd = pd.Series([1,3,5,7,9])

res = pd.concat([even, odd])
print(res)
```

```
0     2
1     4
2     6
3     8
4    10
0     1
1     3
2     5
3     7
4     9
dtype: int64
```

In [23]:
```
res.reset_index(drop=True)
```

Out[23]:
```
0     2
1     4
2     6
3     8
4    10
5     1
6     3
7     5
8     7
9     9
dtype: int64
```

# Sorting

In [24]:
```
 #sort values based on "AveragePrice" (ascending) and "year" (descending)
avocado.sort_values(["AveragePrice", "year"], ascending=[True, False])
```

Out[24]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | |
|---|---|---|---|---|---|---|---|---|---|
| **15261** | 43 | 05-03-2017 | 0.44 | 64057.04 | 223.84 | 4748.88 | 0.00 | 59084.32 | |
| **7412** | 47 | 05-02-2017 | 0.46 | 2200550.27 | 1200632.86 | 531226.65 | 18324.93 | 450365.83 | 11 |
| **15473** | 43 | 05-03-2017 | 0.48 | 50890.73 | 717.57 | 4138.84 | 0.00 | 46034.32 | |
| **15262** | 44 | 26-02-2017 | 0.49 | 44024.03 | 252.79 | 4472.68 | 0.00 | 39298.56 | |
| **1716** | 0 | 27-12-2015 | 0.49 | 1137707.43 | 738314.80 | 286858.37 | 11642.46 | 100891.80 | 7 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **16720** | 18 | 27-08-2017 | 3.04 | 12656.32 | 419.06 | 4851.90 | 145.09 | 7240.27 | |
| **16055** | 42 | 12-03-2017 | 3.05 | 2068.26 | 1043.83 | 77.36 | 0.00 | 947.07 | |
| **14124** | 7 | 06-11-2016 | 3.12 | 19043.80 | 5898.49 | 10039.34 | 0.00 | 3105.97 | |
| **17428** | 37 | 16-04-2017 | 3.17 | 3018.56 | 1255.55 | 82.31 | 0.00 | 1680.70 | |
| **14125** | 8 | 30-10-2016 | 3.25 | 16700.94 | 2325.93 | 11142.85 | 0.00 | 3232.16 | |

18249 rows × 14 columns

# Subsetting

In [25]:
```
# Subsetting columns
avocado["AveragePrice"]
```

Out[25]:
```
0        1.33
1        1.35
2        0.93
3        1.08
4        1.28
         ...
18244    1.63
18245    1.71
18246    1.87
18247    1.93
18248    1.62
Name: AveragePrice, Length: 18249, dtype: float64
```

# Subsetting Multiple Columns

In [26]:
```
# Subsetting multiple columns
avocado[["AveragePrice","Date"]]
```

Out[26]:

| | AveragePrice | Date |
|---|---|---|
| **0** | 1.33 | 27-12-2015 |
| **1** | 1.35 | 20-12-2015 |
| **2** | 0.93 | 13-12-2015 |
| **3** | 1.08 | 06-12-2015 |
| **4** | 1.28 | 29-11-2015 |
| **...** | ... | ... |
| **18244** | 1.63 | 04-02-2018 |
| **18245** | 1.71 | 28-01-2018 |
| **18246** | 1.87 | 21-01-2018 |
| **18247** | 1.93 | 14-01-2018 |
| **18248** | 1.62 | 07-01-2018 |

18249 rows × 2 columns

# Subsetting Rows

In [27]:
```
# Subsetting rows
avocado["AveragePrice"]<1
```

Out[27]:
```
0        False
1        False
2         True
3        False
4        False
         ...
18244    False
18245    False
18246    False
18247    False
18248    False
Name: AveragePrice, Length: 18249, dtype: bool
```

In [28]:
```
# This will print only the rows with price < 1
avocado[avocado["AveragePrice"]<1]
```

Out[28]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Sma Bag |
|---|---|---|---|---|---|---|---|---|---|
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.2 |
| **6** | 6 | 15-11-2015 | 0.99 | 83453.76 | 1368.92 | 73672.72 | 93.26 | 8318.86 | 8196.8 |
| **7** | 7 | 08-11-2015 | 0.98 | 109428.33 | 703.75 | 101815.36 | 80.00 | 6829.22 | 6266.8 |
| **13** | 13 | 27-09-2015 | 0.99 | 106803.39 | 1204.88 | 99409.21 | 154.84 | 6034.46 | 5888.8 |
| **43** | 43 | 01-03-2015 | 0.99 | 55595.74 | 629.46 | 45633.34 | 181.49 | 9151.45 | 8986.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **17169** | 43 | 05-03-2017 | 0.99 | 155011.12 | 35367.23 | 5175.81 | 5.91 | 114462.17 | 95379.0 |
| **17170** | 44 | 26-02-2017 | 0.99 | 171145.00 | 34520.03 | 6936.39 | 0.00 | 129688.58 | 117252.3 |
| **17536** | 39 | 02-04-2017 | 0.98 | 402676.23 | 34093.33 | 58330.53 | 207.85 | 310044.52 | 155701.4 |
| **17537** | 40 | 26-03-2017 | 0.90 | 456645.91 | 36169.35 | 51398.72 | 139.55 | 368938.29 | 152159.5 |
| **17540** | 43 | 05-03-2017 | 0.99 | 367519.17 | 61166.48 | 55123.99 | 126.80 | 251101.90 | 112844.1 |

2796 rows × 14 columns

# Subsetting Based on Data

In [29]:
```
avocado[avocado["Date"]<="2015-02-04"]
```

Out[29]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 |
| **6** | 6 | 15-11-2015 | 0.99 | 83453.76 | 1368.92 | 73672.72 | 93.26 | 8318.86 | 8196.81 |
| **7** | 7 | 08-11-2015 | 0.98 | 109428.33 | 703.75 | 101815.36 | 80.00 | 6829.22 | 6266.85 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18242** | 5 | 18-02-2018 | 1.56 | 17597.12 | 1892.05 | 1928.36 | 0.00 | 13776.71 | 13553.53 |
| **18243** | 6 | 11-02-2018 | 1.57 | 15986.17 | 1924.28 | 1368.32 | 0.00 | 12693.57 | 12437.35 |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 |

12094 rows × 14 columns

# Subsetting based on Multiple Conditions

In [30]:
```
avocado[(avocado["Date"]<"2015-02-04") & (avocado["type"]=="organic")]
```

Out[30]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | La B |
|---|---|---|---|---|---|---|---|---|---|---|
| **9127** | 1 | 20-12-2015 | 1.89 | 1163.03 | 30.24 | 172.14 | 0.00 | 960.65 | 960.65 | 0 |
| **9128** | 2 | 13-12-2015 | 1.85 | 995.96 | 10.44 | 178.70 | 0.00 | 806.82 | 806.82 | 0 |
| **9129** | 3 | 06-12-2015 | 1.84 | 1158.42 | 90.29 | 104.18 | 0.00 | 963.95 | 948.52 | 15 |
| **9132** | 6 | 15-11-2015 | 1.89 | 1208.54 | 20.71 | 238.16 | 0.00 | 949.67 | 949.67 | 0 |
| **9133** | 7 | 08-11-2015 | 1.88 | 1332.27 | 20.08 | 351.40 | 0.00 | 960.79 | 960.79 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **18242** | 5 | 18-02-2018 | 1.56 | 17597.12 | 1892.05 | 1928.36 | 0.00 | 13776.71 | 13553.53 | 223 |
| **18243** | 6 | 11-02-2018 | 1.57 | 15986.17 | 1924.28 | 1368.32 | 0.00 | 12693.57 | 12437.35 | 256 |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 | 431 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 | 50 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 | 26 |

6046 rows × 14 columns

# Subsetting using . isin()

In [31]:
```
regionFilter = avocado["region"].isin(["Boston", "SanDiego"])
avocado[regionFilter]
```

Out[31]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Smal Bags |
|---|---|---|---|---|---|---|---|---|---|
| **208** | 0 | 27-12-2015 | 1.13 | 450816.39 | 3886.27 | 346964.70 | 13952.56 | 86012.86 | 85913.6( |
| **209** | 1 | 20-12-2015 | 1.07 | 489802.88 | 4912.37 | 390100.99 | 5887.72 | 88901.80 | 88768.47 |
| **210** | 2 | 13-12-2015 | 1.01 | 549945.76 | 4641.02 | 455362.38 | 219.40 | 89722.96 | 89523.38 |
| **211** | 3 | 06-12-2015 | 1.02 | 488679.31 | 5126.32 | 407520.22 | 142.99 | 75889.78 | 75666.22 |
| **212** | 4 | 29-11-2015 | 1.19 | 350559.81 | 3609.25 | 272719.08 | 105.86 | 74125.62 | 73864.52 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18100** | 7 | 04-02-2018 | 1.81 | 17454.74 | 1158.41 | 7388.27 | 0.00 | 8908.06 | 8908.06 |
| **18101** | 8 | 28-01-2018 | 1.91 | 17579.47 | 1145.64 | 8284.41 | 0.00 | 8149.42 | 8149.42 |
| **18102** | 9 | 21-01-2018 | 1.95 | 18676.37 | 1088.49 | 9282.37 | 0.00 | 8305.51 | 8305.51 |
| **18103** | 10 | 14-01-2018 | 1.81 | 21770.02 | 3285.98 | 14338.52 | 0.00 | 4145.52 | 4145.52 |
| **18104** | 11 | 07-01-2018 | 2.06 | 16746.82 | 5150.82 | 9366.31 | 0.00 | 2229.69 | 2229.69 |

676 rows × 14 columns

# Multiple Parameter Filtering

In [32]:
```
regionFilter = avocado["region"].isin(["Boston", "SanDiego"])
yearFilter = avocado["year"].isin(["2016", "2017"])
avocado[regionFilter & yearFilter]
```

Out[32]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags | XLarge Bags | type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

# Detecting Missing Values.isna()

In [33]:
```
avocado.isna()
```

Out[33]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags | XLarge Bags |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False | False | False | False | False |
| **1** | False | False | False | False | False | False | False | False | False | False | False |
| **2** | False | False | False | False | False | False | False | False | False | False | False |
| **3** | False | False | False | False | False | False | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False | False | False | False | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18244** | False | False | False | False | False | False | False | False | False | False | False |
| **18245** | False | False | False | False | False | False | False | False | False | False | False |
| **18246** | False | False | False | False | False | False | False | False | False | False | False |
| **18247** | False | False | False | False | False | False | False | False | False | False | False |
| **18248** | False | False | False | False | False | False | False | False | False | False | False |

18249 rows × 14 columns

In [34]:
```
avocado.isna().any()
```

Out[34]:
```
Unnamed: 0      False
Date            False
AveragePrice    False
Total Volume    False
4046            False
4225            False
4770            False
Total Bags      False
Small Bags      False
Large Bags      False
XLarge Bags     False
type            False
year            False
region          False
dtype: bool
```

# Counting missing Values

In [35]:
```
avocado.isna().sum()
```

Out[35]:
```
Unnamed: 0      0
Date            0
AveragePrice    0
Total Volume    0
4046            0
4225            0
4770            0
Total Bags      0
Small Bags      0
Large Bags      0
XLarge Bags     0
type            0
year            0
region          0
dtype: int64
```

# Removing Missing Values

In [36]:
```
avocado.dropna()
meanVal = avocado["AveragePrice"].mean()
avocado.fillna(meanVal)
```

Out[36]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 |
| **18245** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.84 | 8940.04 |
| **18246** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.11 | 9351.80 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 |

18249 rows × 14 columns

# Adding a new column

In [37]:
```
avocado["AveragePricePer100"] = avocado["AveragePrice"] * 100
avocado
```

Out[37]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 |
| **18245** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.84 | 8940.04 |
| **18246** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.11 | 9351.80 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 |

18249 rows × 15 columns

# Delecting Columns in DataFrame

# .drop(lst,axis=1)

In [38]:
```
avocado.drop(["AveragePricePer100"],axis = 1)
```

Out[38]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 |
| **18245** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.84 | 8940.04 |
| **18246** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.11 | 9351.80 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 |

18249 rows × 14 columns

# Summary Statistics

In [39]:
```
# mean of the AveragePrice of avocado
avocado["AveragePrice"].mean()
```

Out[39]:
```
1.405978409775878
```

# Sumarizing Dates

In [40]:
```
avocado["Date"].max()
```

Out[40]:
```
'31-12-2017'
```

# .agg() method

In [41]:
```
def pct30(column):
    #return the 0.3 quartile
    return column.quantile(0.3)
def pct50(column):
    #return the 0.5 quartile
    return column.quantile(0.5)

avocado[["AveragePrice","Total Bags"]].agg([pct30,pct50])
```

Out[41]:

| | AveragePrice | Total Bags |
|---|---|---|
| **pct30** | 1.15 | 7316.634 |
| **pct50** | 1.37 | 39743.830 |

# Dropping duplicate names

# .drop_duplicates(lst)

Delete all the duplicate names from the dataframe

In [43]:
```
temp = avocado.drop_duplicates(subset=["year"])
temp
```

Out[43]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 |
| **2808** | 0 | 25-12-2016 | 1.52 | 73341.73 | 3202.39 | 58280.33 | 426.92 | 11432.09 | 11017.32 |
| **5616** | 0 | 31-12-2017 | 1.47 | 113514.42 | 2622.70 | 101135.53 | 20.25 | 9735.94 | 5556.98 |
| **8478** | 0 | 25-03-2018 | 1.57 | 149396.50 | 16361.69 | 109045.03 | 65.45 | 23924.33 | 19273.80 |

# Count Categorical data .value_counts()

In [44]:
```
# count number of avocado in each year in descending order
avocado["year"].value_counts(sort=True, ascending = False)
```

Out[44]:
```
year
2017    5722
2016    5616
2015    5615
2018    1296
Name: count, dtype: int64
```

# Grouped Summaries.groupby(col)

In [45]:
```
# group by multiple columns and perform multiple summary statistic operations
avocado.groupby(["year","type"])["AveragePrice"].agg([min,max,np.mean,np.median])
```

Out[45]:

| | | min | max | mean | median |
|---|---|---|---|---|---|
| **year** | **type** | | | | |
| **2015** | **conventional** | 0.49 | 1.59 | 1.077963 | 1.08 |
| | **organic** | 0.81 | 2.79 | 1.673324 | 1.67 |
| **2016** | **conventional** | 0.51 | 2.20 | 1.105595 | 1.08 |
| | **organic** | 0.58 | 3.25 | 1.571684 | 1.53 |
| **2017** | **conventional** | 0.46 | 2.22 | 1.294888 | 1.30 |
| | **organic** | 0.44 | 3.17 | 1.735521 | 1.72 |
| **2018** | **conventional** | 0.56 | 1.74 | 1.127886 | 1.14 |
| | **organic** | 1.01 | 2.30 | 1.567176 | 1.55 |

# Pivot Table

In [46]:
```
# this is the same table we build in the previous cell but using pivot table
avocado.pivot_table(index=["year","type"], aggfunc=[min,max,np.mean,np.median], val
```

Out[46]:

| | | min | max | mean | median |
|---|---|---|---|---|---|
| | | **AveragePrice** | **AveragePrice** | **AveragePrice** | **AveragePrice** |
| **year** | **type** | | | | |
| **2015** | **conventional** | 0.49 | 1.59 | 1.077963 | 1.08 |
| | **organic** | 0.81 | 2.79 | 1.673324 | 1.67 |
| **2016** | **conventional** | 0.51 | 2.20 | 1.105595 | 1.08 |
| | **organic** | 0.58 | 3.25 | 1.571684 | 1.53 |
| **2017** | **conventional** | 0.46 | 2.22 | 1.294888 | 1.30 |
| | **organic** | 0.44 | 3.17 | 1.735521 | 1.72 |
| **2018** | **conventional** | 0.56 | 1.74 | 1.127886 | 1.14 |
| | **organic** | 1.01 | 2.30 | 1.567176 | 1.55 |

# Explicit Indexes

In [47]:
```
regionIndex = avocado.set_index(["region"])
regionIndex
```

Out[47]:

| region | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Tot Bag |
|---|---|---|---|---|---|---|---|---|
| **Albany** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.8 |
| **Albany** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.5 |
| **Albany** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.3 |
| **Albany** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.1 |
| **Albany** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.9 |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **WestTexNewMexico** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.6 |
| **WestTexNewMexico** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.8 |
| **WestTexNewMexico** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.1 |
| **WestTexNewMexico** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.5 |
| **WestTexNewMexico** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.1 |

18249 rows × 14 columns

In [48]:
```
# Insted of doing this
avocado[avocado["region"].isin(["Albany", "WestTexNewMexico"])]
```

Out[48]:

| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.87 | 8603.62 |
| **1** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.56 | 9408.07 |
| **2** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.35 | 8042.21 |
| **3** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.16 | 5677.40 |
| **4** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.95 | 5986.26 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **18244** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.67 | 13066.82 |
| **18245** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.84 | 8940.04 |
| **18246** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.11 | 9351.80 |
| **18247** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.54 | 10919.54 |
| **18248** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.15 | 11988.14 |

673 rows × 15 columns

In [49]:
```
# we can simply do
regionIndex.loc[["Albany", "WestTexNewMexico"]]
```

Out[49]:

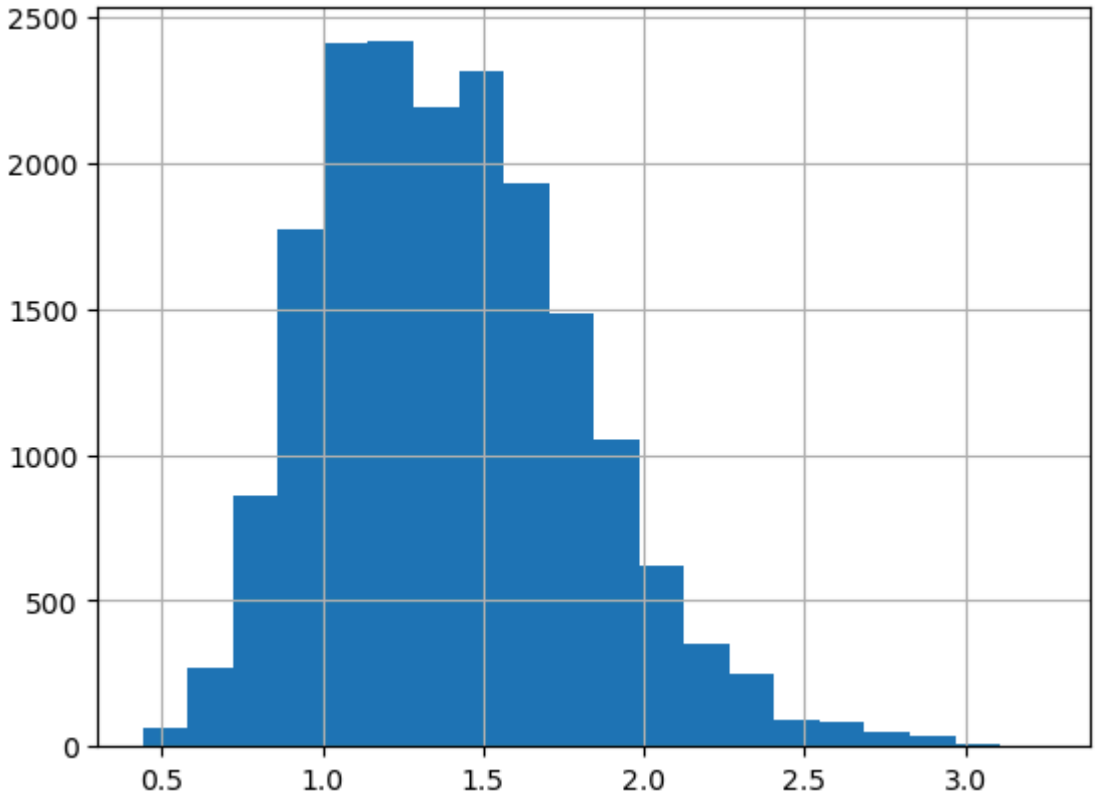
| region | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Tot Bag |
|---|---|---|---|---|---|---|---|---|
| **Albany** | 0 | 27-12-2015 | 1.33 | 64236.62 | 1036.74 | 54454.85 | 48.16 | 8696.8 |
| **Albany** | 1 | 20-12-2015 | 1.35 | 54876.98 | 674.28 | 44638.81 | 58.33 | 9505.5 |
| **Albany** | 2 | 13-12-2015 | 0.93 | 118220.22 | 794.70 | 109149.67 | 130.50 | 8145.3 |
| **Albany** | 3 | 06-12-2015 | 1.08 | 78992.15 | 1132.00 | 71976.41 | 72.58 | 5811.1 |
| **Albany** | 4 | 29-11-2015 | 1.28 | 51039.60 | 941.48 | 43838.39 | 75.78 | 6183.9 |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **WestTexNewMexico** | 7 | 04-02-2018 | 1.63 | 17074.83 | 2046.96 | 1529.20 | 0.00 | 13498.6 |
| **WestTexNewMexico** | 8 | 28-01-2018 | 1.71 | 13888.04 | 1191.70 | 3431.50 | 0.00 | 9264.8 |
| **WestTexNewMexico** | 9 | 21-01-2018 | 1.87 | 13766.76 | 1191.92 | 2452.79 | 727.94 | 9394.1 |
| **WestTexNewMexico** | 10 | 14-01-2018 | 1.93 | 16205.22 | 1527.63 | 2981.04 | 727.01 | 10969.5 |
| **WestTexNewMexico** | 11 | 07-01-2018 | 1.62 | 17489.58 | 2894.77 | 2356.13 | 224.53 | 12014.1 |

673 rows × 14 columns

# Visualizing Data

# Histogram

In [54]:
```
avocado["AveragePrice"].hist(bins=20)
plt.show()
```



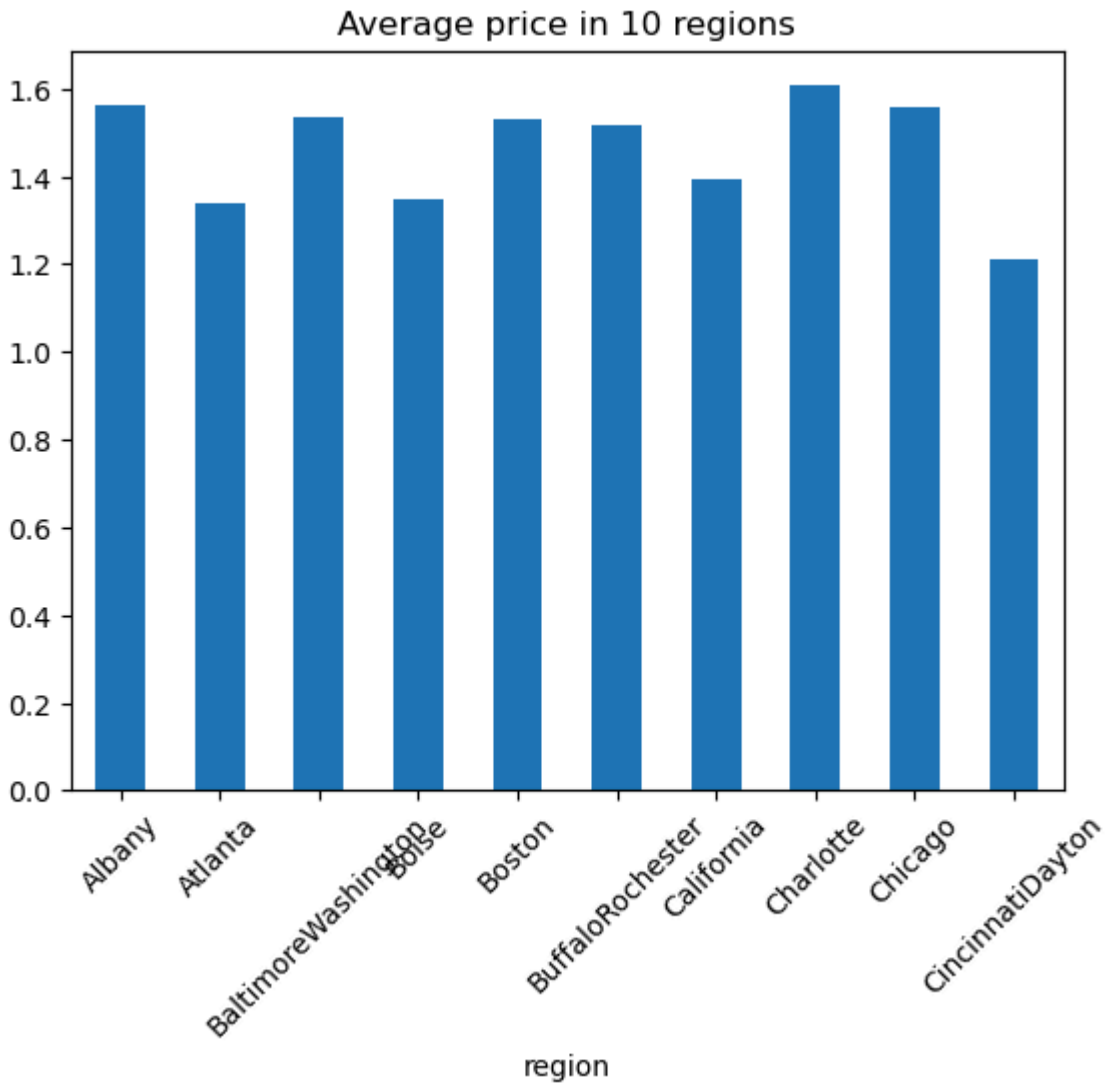

In [51]:
```
regionFilter = avocado.groupby("region")["AveragePrice"].mean().head(10)
regionFilter
```

Out[51]:
```
region
Albany                  1.561036
Atlanta                 1.337959
BaltimoreWashington     1.534231
Boise                   1.348136
Boston                  1.530888
BuffaloRochester        1.516834
California              1.395325
Charlotte               1.606036
Chicago                 1.556775
CincinnatiDayton        1.209201
Name: AveragePrice, dtype: float64
```

# Bar plot

In [52]:
```
regionFilter.plot(kind = "bar",rot=45,title="Average price in 10 regions")
```

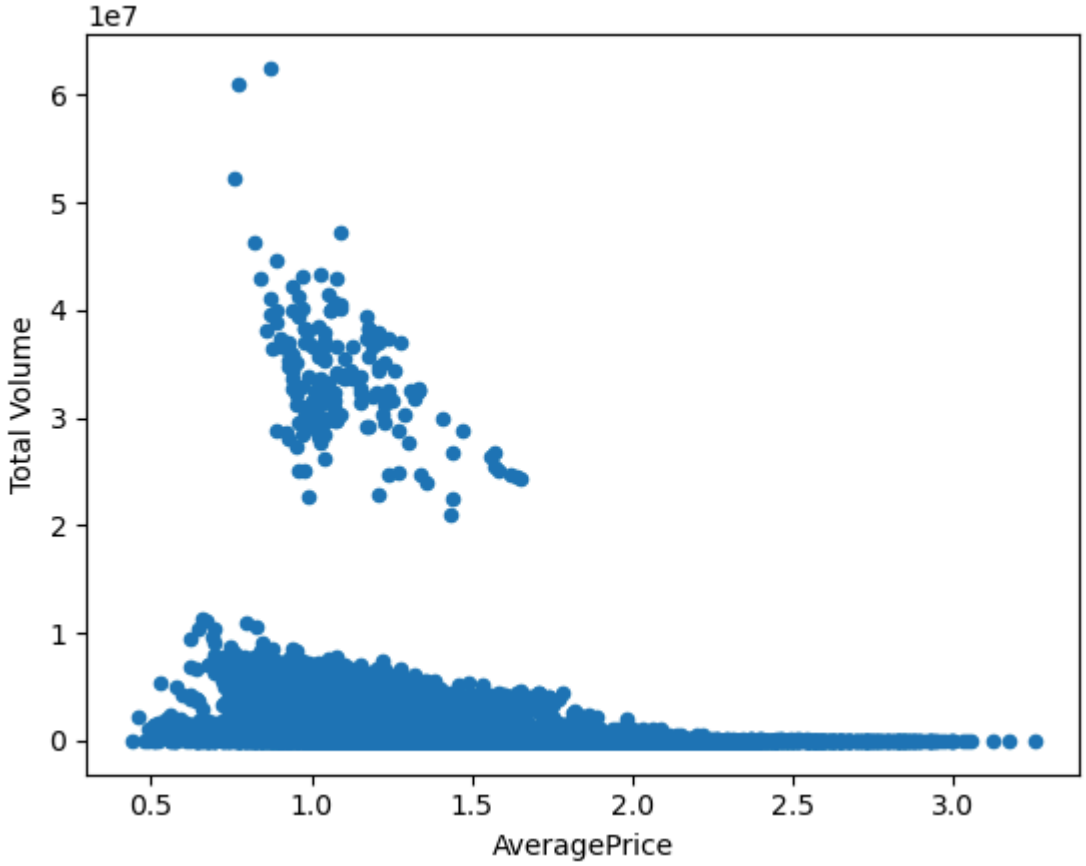
Out[52]:
```
<Axes: title={'center': 'Average price in 10 regions'}, xlabel='region'>
```




# #Scatter Plot

In [55]:
```
avocado.plot(x="AveragePrice", y="Total Volume", kind="scatter")
```

Out[55]:
```
<Axes: xlabel='AveragePrice', ylabel='Total Volume'>
```




In [ ]: