

```
import pandas as pd
from pandas.plotting import scatter_matrix
import numpy as np
from numpy import percentile
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
from sklearn import preprocessing
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.preprocessing import StandardScaler
import sklearn.metrics as metrics
import statsmodels.api as sm
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

In [1]:
f1=pd.read_csv("avocado.csv.zip")

Out[2]:

```

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
...
18244	7	2018-02-04	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	2018	WestTexNewMexico
18245	8	2018-01-28	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	2018	WestTexNewMexico
18246	9	2018-01-21	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	2018	WestTexNewMexico
18247	10	2018-01-14	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	2018	WestTexNewMexico
18248	11	2018-01-07	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	2018	WestTexNewMexico

18249 rows x 14 columns

```
In [3]:
f1.tail()
```

```
Out[3]:

```

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
18244	7	2018-02-04	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	2018	WestTexNewMexico
18245	8	2018-01-28	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	2018	WestTexNewMexico
18246	9	2018-01-21	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	2018	WestTexNewMexico
18247	10	2018-01-14	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	2018	WestTexNewMexico
18248	11	2018-01-07	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	2018	WestTexNewMexico

```
In [4]:
f1.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: 1.8+ MB
```

```
In [5]:
f1.describe()
```

```


```

	Unnamed: 0	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	year
count	18249.000000	18249.000000	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	1.824900e+04	18249.000000	18249.000000
mean	24.232232	1.405978	8.506440e+05	2.930084e+05	2.951546e+05	2.283974e+04	2.396392e+05	1.821947e+05	5.433809e+04	3106.426507	2016.147899
std	15.481045	0.402977	3.453545e+06	1.264989e+06	1.204120e+06	1.074641e+05	9.862424e+05	7.461785e+05	2.439060e+05	17882.884652	0.939638
min	0.000000	0.440050	8.456002e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	2015.000000
25%	151500000	1.100000	1.063858e+04	8.540770e+02	3.008780e+03	0.000000e+00	5.088640e+03	2.848402e+03	1.274710e+02	0.000000	2015.000000
50%	24.000000	1.370000	1.073780e+06	8.645300e+03	2.906102e+04	1.846990e+02	3.374383e+04	2.836282e+04	2.647710e+03	0.000000	2016.000000
75%	38.000000	1.660000	4.329632e+05	1.110200e+05	1.502069e+05	6.243420e+03	1.107834e+05	6.333767e+04	2.202325e+04	132.500000	2017.000000
max	52.000000	3.250000	6.250565e+07	2.274362e+07	2.047057e+07	2.546439e+06	1.937313e+07	1.338459e+07	5.719076e+06	551693.650000	2018.000000

```
In [6]:
f1.shape
```

```
Out[6]:
(18249, 14)
```

```
In [7]:
f1.AveragePrice.plot(kind='hist', figsize=(10,10), color="red",)
plt.show()
```

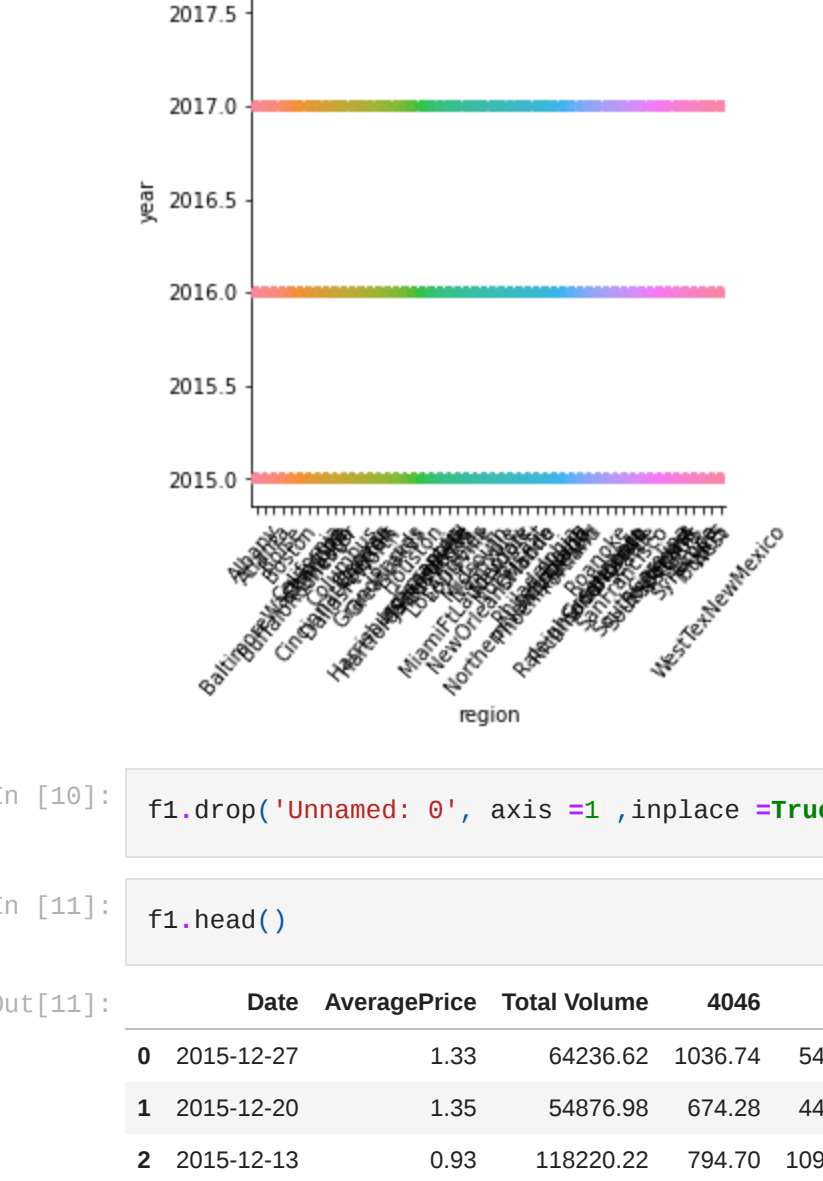


```
In [8]:
scatter_matrix(f1 , figsize=(15,15) ,diagonal="hist" ,color="r")
plt.show()
```



```
In [9]:
plt.figure(figsize=(50,50))
sns.catplot(x='region', y='year', data=f1)
plt.xticks(rotation=50)
plt.show()
```

<Figure size 3600x3600 with 0 Axes>



```
In [10]:
f1.drop('Unnamed: 0', axis =1 ,inplace=True)
```

```
In [11]:
f1.head()
```

```
Out[11]:

```

	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

```
In [12]:
plt.figure(figsize=(10,10),dpi=90)
sns.boxplot(data = f1[[
'AveragePrice',
'Total Volume',
'4046',
'4225',
'4770',
'Total Bags',
'Small Bags',
'Large Bags',
'XLarge Bags',
'type'
]])
plt.show()
```



```
In [13]:
f1.drop(columns=['Date'],inplace=True)
f1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18249 entries, 0 to 18248
Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   AveragePrice         18249 non-null  float64
1   Total Volume         18249 non-null  float64
2   4046                 18249 non-null  float64
3   4225                 18249 non-null  float64
4   4770                 18249 non-null  float64
5   Total Bags           18249 non-null  float64
6   Small Bags           18249 non-null  float64
7   Large Bags           18249 non-null  float64
8   XLarge Bags          18249 non-null  float64
9   type                 18249 non-null  object
10  year                 18249 non-null  int64
11  region               18249 non-null  object
dtypes: float64(9), int64(1), object(2)
memory usage: 1.7+ MB
```

```
In [14]:
columns = f1.columns
for i in columns:
    if isinstance(f1[i][0], str) :
        else continue
    #defining quartiles
    quartiles = percentile(f1[i], [20,60])
    # calculate min/max
    lower_fence = quartiles[0] - (1.5*(quartiles[1]-quartiles[0]))
    upper_fence = quartiles[1] + (1.5*(quartiles[2]-quartiles[0]))
    f1[i] = f1[i].apply(lambda x: upper_fence if x > upper_fence else (lower_fence if x < lower_fence else x))
```

```
In [15]:
f1
```

```
Out[15]:

```

	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015.0	Albany
1	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015.0	Albany
2	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015.0	Albany
3	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015.0	Albany
4	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015.0	Albany
...
18244	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	2017.5	WestTexNewMexico
18245	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	2017.5	WestTexNewMexico
18246	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	2017.5	WestTexNewMexico
18247	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	2017.5	WestTexNewMexico
18248	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	2017.5	WestTexNewMexico

18249 rows x 12 columns

```
In [16]:
plt.figure(figsize=(10,5),dpi=80)
sns.catplot(data = f1[[
'AveragePrice',
'Total Volume',
'4046',
'4225',
'4770',
'Total Bags',
'Small Bags',
'Large Bags',
'XLarge Bags',
'type'
]])
plt.show()
```

<Figure size 800x400 with 0 Axes>



```
In [20]:
f1['region'] = pd.Categorical(f1['region'])
dummies = pd.get_dummies(f1['region'], prefix = 'region')
dummies
```

```
Out[20]:

```

	region_Albany	region_Antiochia	region_BatimoreWashington	region_Boise	region_Boston	region_BuffaloRochester	region_California	region_Charlotte	region_Chicago	region_CincinnatiDayton	region_SouthCarolina	region_South
0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
...
18244	0	0	0	0	0	0	0	0	0	0	0	0
18245	0	0	0	0	0	0	0	0	0	0	0	0
18246	0	0	0	0	0	0	0	0	0	0	0	0
18247	0	0	0	0	0	0	0	0	0	0	0	0
18248	0	0	0	0	0	0	0	0	0	0	0	0

18249 rows x 54 columns

```
In [22]:
data = pd.concat([f1, dummies], axis=1)
data.drop(columns="region",inplace=True)
data
```

```
Out[22]:

```

	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	region_SouthCarolina	region_SouthCentral	region_Southeast	region_Spokane	region_StLouis	region_Syracuse	region_Tampa	region_Texas
0	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	0	0	0	0	0	0	0	0
1	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	0	0	0	0	0	0	0	0
2	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	0	0	0	0	0	0	0	0
3	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	0	0	0	0	0	0	0	0
4	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	0	0	0	0	0	0	0	0
...
18244	1.63	17074.83	2046.96	1529.20	0.00	13498.67	13066.82	431.85	0.0	organic	0	0	0	0	0	0	0	0
18245	1.71	13888.04	1191.70	3431.50	0.00	9264.84	8940.04	324.80	0.0	organic	0	0	0	0	0	0	0	0
18246	1.87	13766.76	1191.92	2452.79	727.94	9394.11	9351.80	42.31	0.0	organic	0	0	0	0	0	0	0	0
18247	1.93	16205.22	1527.63	2981.04	727.01	10969.54	10919.54	50.00	0.0	organic	0	0	0	0	0	0	0	0
18248	1.62	17489.58	2894.77	2356.13	224.53	12014.15	11988.14	26.01	0.0	organic	0	0	0	0	0	0	0	0

18249 rows x 65 columns

```
In [24]:
# label encoding
label_encoder = preprocessing.LabelEncoder()
data['type'] = label_encoder.fit_transform(data['type'])
data
```

```
Out[24]:

```

	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	region_SouthCarolina	region_SouthCentral	region_Southeast	region_Spokane	region_StLouis	region_Syracuse	region_Tampa	region_Texas
0	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	0	0	0	0	0	0	0	0	0
1	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	0	0	0	0	0	0	0	0	0
2	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	0	0	0	0	0	0	0	0	0
3	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	13										