Cargando las librerias y el dataset a utilizar

In [56]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [57]:
```
df = pd.read_csv('avocado_clean.csv')
```

Medidas de tendencia central (Media, Mediana y Moda), Medidas de dispersión(varianza, desviación estándar, rango intercuartílico) y análisis de resultados

Medidas de tendencia central

In [58]:
```
df.head()
```

Out[58]:

	Date	AveragePrice	Total Volume	4046	4225	Total Bags	Small Bags	
0	2015-12-27	1.33	64236.62	1036.74	54454.85	8696.87	8603.62	conve
1	2015-12-20	1.35	54876.98	674.28	44638.81	9505.56	9408.07	conve
2	2015-12-13	0.93	118220.22	794.70	109149.67	8145.35	8042.21	conve
3	2015-12-06	1.08	78992.15	1132.00	71976.41	5811.16	5677.40	conve
4	2015-11-29	1.28	51039.60	941.48	43838.39	6183.95	5986.26	conve

In [59]:
```
#Media

numerical_columns = df.select_dtypes(include='number').mean()

mean_values = numerical_columns.to_dict()

print("\nMedia\n")
for key, value in mean_values.items():
    print(f"{key}: {value}")
```

```python
#Mediana

numerical_columns = df.select_dtypes(include='number').median()

median_values = numerical_columns.to_dict()

print("\nMediana\n")
for key, value in median_values.items():
    print(f"{key}: {value}")

#Mode

mode_values = df.mode().loc[:,  ['type', 'region']]

print("\nModa\n")
for key, value in mode_values.items():
    print(f"{key}: {value[0]}")
    print(f"{key}: {value[1]}")

    print(f'veces repetidas: {df[key].value_counts()[value[0]]}')

    print('')
```

```
Media

AveragePrice: 1.5335278538812782
Total Volume: 51930.60923287671
4046: 13106.127513242007
4225: 18230.428082191782
Total Bags: 18907.94536073059
Small Bags: 13440.854275799085

Mediana

AveragePrice: 1.52
Total Volume: 16174.650000000001
4046: 1487.73
4225: 4496.299999999999
Total Bags: 7777.965
Small Bags: 5263.625

Moda

type: organic
type: nan
veces repetidas: 8304

region: Louisville
region: Syracuse
veces repetidas: 338
```

medidas de dispersión

In [60]:

```python
#Varianza

numerical_columns = df.select_dtypes(include='number').var()

var_values = numerical_columns.to_dict()

print("\nVarianza\n")
for key, value in var_values.items():
    print(f"{key}: {value}")


#Desviación estandar

print("\nDesviación estandar\n")
for key, value in var_values.items():
    print(f"{key}: {value**(1/2)}")


#rango intercuartílico

numerical_columns = df.select_dtypes(include='number')

Q1 = numerical_columns.quantile(0.25)
Q3 = numerical_columns.quantile(0.75)
IQR = Q3 - Q1

print("\nRango intercuartílico\n")
for key, value in IQR.items():
    print(f"{key}: {value}")
```

```
Varianza

AveragePrice: 0.12974655393228388
Total Volume: 4582538065.524958
4046: 747214308.205814
4225: 749621654.2896923
Total Bags: 521105563.58260655
Small Bags: 289719364.5196014

Desviación estandar

AveragePrice: 0.36020348961702725
Total Volume: 67694.44634181565
4046: 27335.22101988228
4225: 27379.219387880516
Total Bags: 22827.73671616629
Small Bags: 17021.14463012407

Rango intercuartílico

AveragePrice: 0.52
Total Volume: 74039.925
4046: 8177.825
4225: 23652.910000000003
Total Bags: 27812.71
Small Bags: 19615.9625
```

Analisis de los resultados

El valor más consistente es el precio promedio del aguacate, ya que tanto su media como mediana tienen casi el mismo valor, además que su desviación estándar es relativamente baja (0.36 dolares).

Donde hay ligeros problemas son en los demás datos numéricos:

la media es mayor a la mediana, indicando que hay varios valores muy grandes que afectan el valor de la media. Esto es más evidente al comparar el rango intercuartílico con la mediana de estas mismas variables: Se observa que es una diferencia bastante grande en comparación de las medianas, indicando que los valores cercanos al percentil 75% empiezan a variar demasiado sus valores con respecto a percentiles más bajos.

También se observa una desviación estandar mayor que la media, a veces hasta el doble de grande como en '4046', indicando una gran dispersión en los datos. Esto sucede para las columnas 'Total Volume', '4046', '4225', 'Total Bags' y 'Small Bags'.

En cuanto a las columnas categóricas:

La columna 'type' muestra como moda el valor 'organic' con 8304 repeticiones. Considerando que el dataset tiene 10,000 filas, saber que la mayoría de

aguacates son orgánicos nos muestra que el dataset está claramente desbalanceado.

Por otro lado, en cuanto a 'region' tanto Louisville y Syracuse son la moda con 338 repeticiones. En este sentido, el dataset probablemente no esté tan desbalanceado, considerando que hay solo 53 regiones diferentes.