Mapas de calor y boxplots

May 11, 2022

1 Actividad Evaluable 3: Mapas de calor y boxplots

1. Carga los datos usando tu lector de csv o con pandas.

[24]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv("avocado.csv")
df.head()
```

[24]:

	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

[43]:
```
df.describe(include = object).transpose()
```

[43]:

	count	unique	top	freq
Date	18249	169	2015-12-27	108
type	18249	2	conventional	9126
region	18249	54	Albany	338

2. Realiza el análisis de las variables usando diagramas de cajas y bigotes, histogramas y mapas de calor.

[44]:
```
counts = df['region'].value_counts()
print(counts)
```

```
Albany                338
Sacramento            338
Northeast             338
NorthernNewEngland    338
Orlando               338
Philadelphia          338
PhoenixTucson         338
Pittsburgh            338
Plains                338
Portland              338
RaleighGreensboro     338
RichmondNorfolk       338
Roanoke               338
SanDiego              338
Atlanta               338
SanFrancisco          338
Seattle               338
SouthCarolina         338
SouthCentral          338
Southeast             338
Spokane               338
StLouis               338
Syracuse              338
Tampa                 338
TotalUS               338
West                  338
NewYork               338
NewOrleansMobile      338
Nashville             338
Midsouth              338
BaltimoreWashington   338
Boise                 338
Boston                338
BuffaloRochester      338
California            338
Charlotte             338
Chicago               338
CincinnatiDayton      338
Columbus              338
DallasFtWorth         338
Denver                338
Detroit               338
```

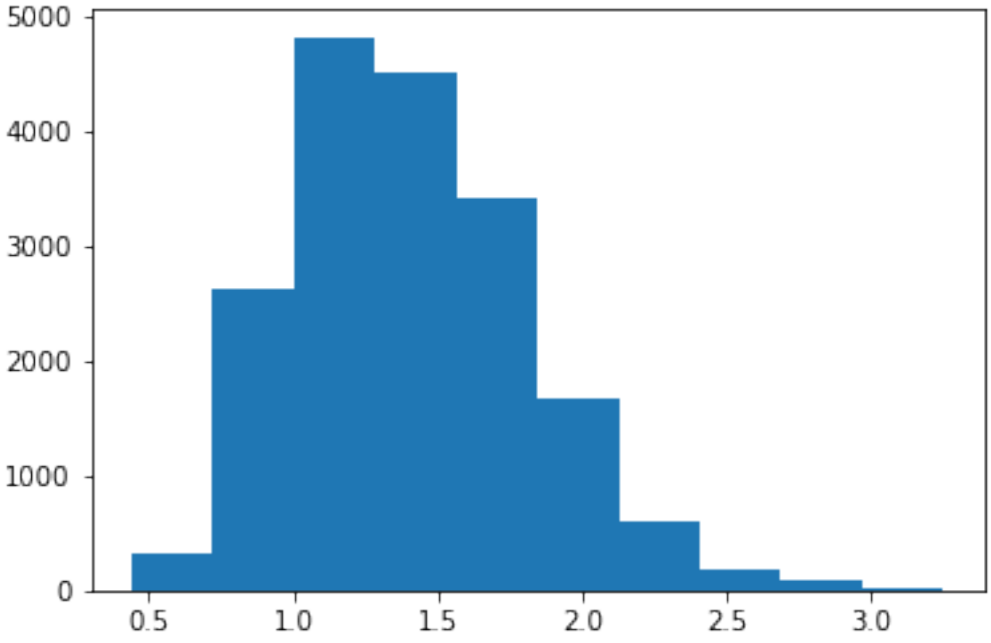
```
GrandRapids             338
GreatLakes              338
HarrisburgScranton      338
HartfordSpringfield     338
Houston                 338
Indianapolis            338
Jacksonville            338
LasVegas                338
LosAngeles              338
Louisville              338
MiamiFtLauderdale       338
WestTexNewMexico        335
Name: region, dtype: int64
```

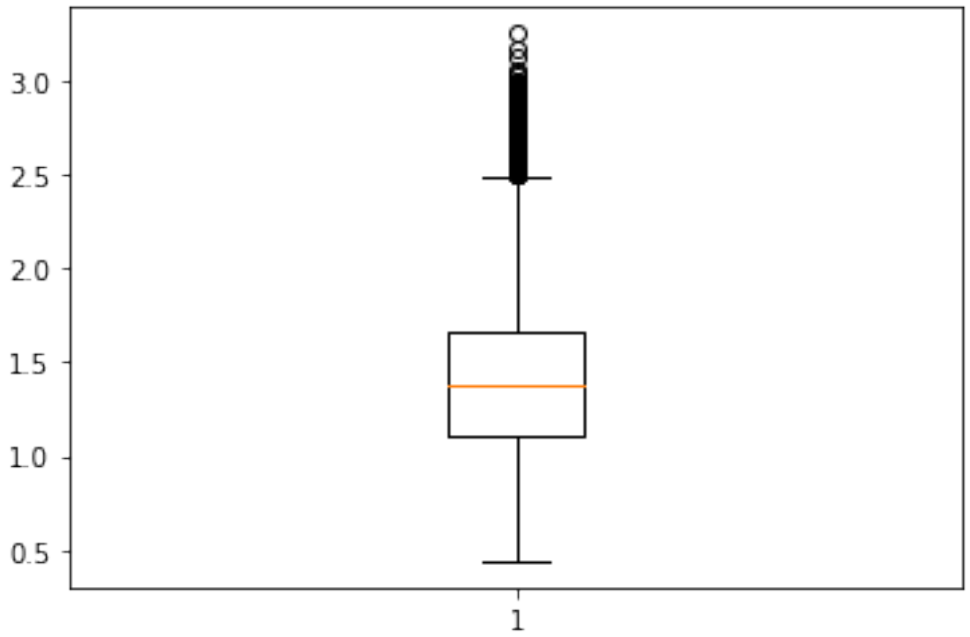
[42]:
```
df.iloc[1:, [2]]
```

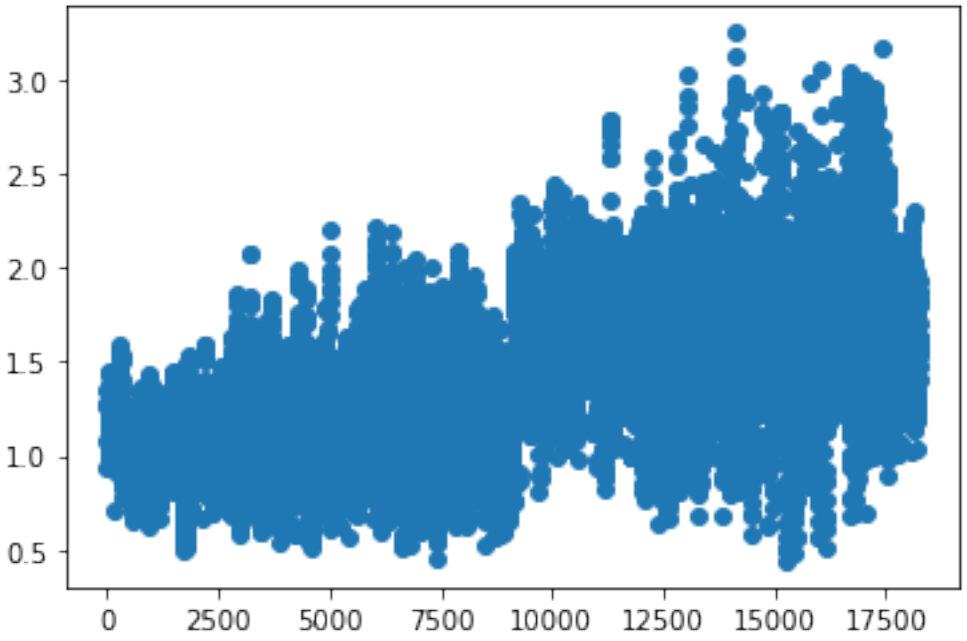
[42]:
```
        AveragePrice
1               1.35
2               0.93
3               1.08
4               1.28
5               1.26
…                …
18244           1.63
18245           1.71
18246           1.87
18247           1.93
18248           1.62

[18248 rows x 1 columns]
```

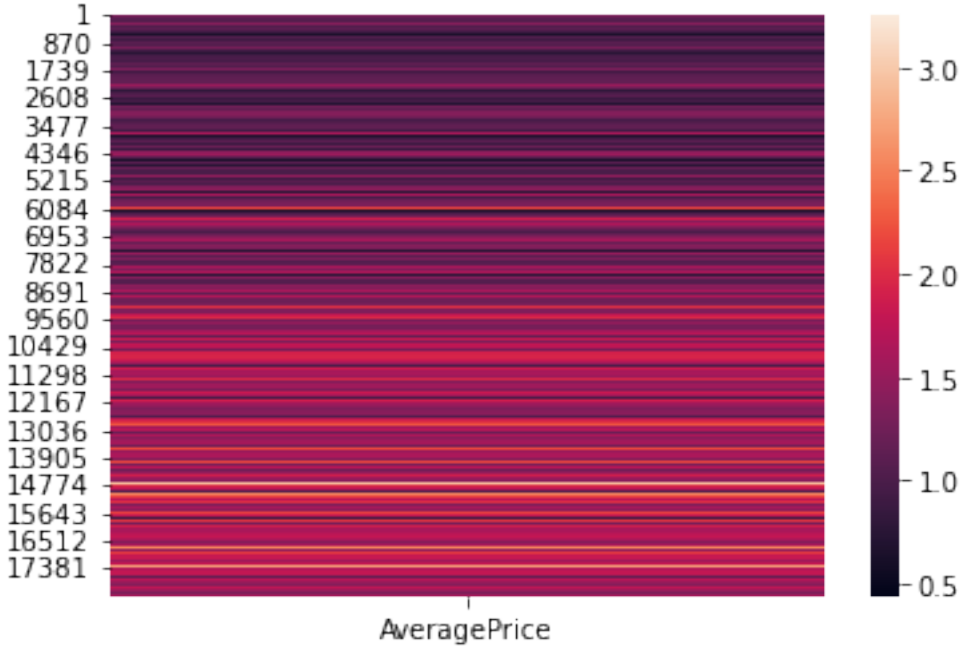
[41]:
```
plt.hist(df.iloc[1:, [2]])
plt.show()
plt.boxplot(df.iloc[1:, [2]])
plt.show()
y=df.iloc[1:, [2]]
x=range(len(y))
plt.scatter(x,y)
plt.show()
```


5000
4000
3000
2000
1000
0
0.5
1.0
1.5
2.0
2.5
3.0


3.0
2.5
2.0
1.5
1.0
0.5
1



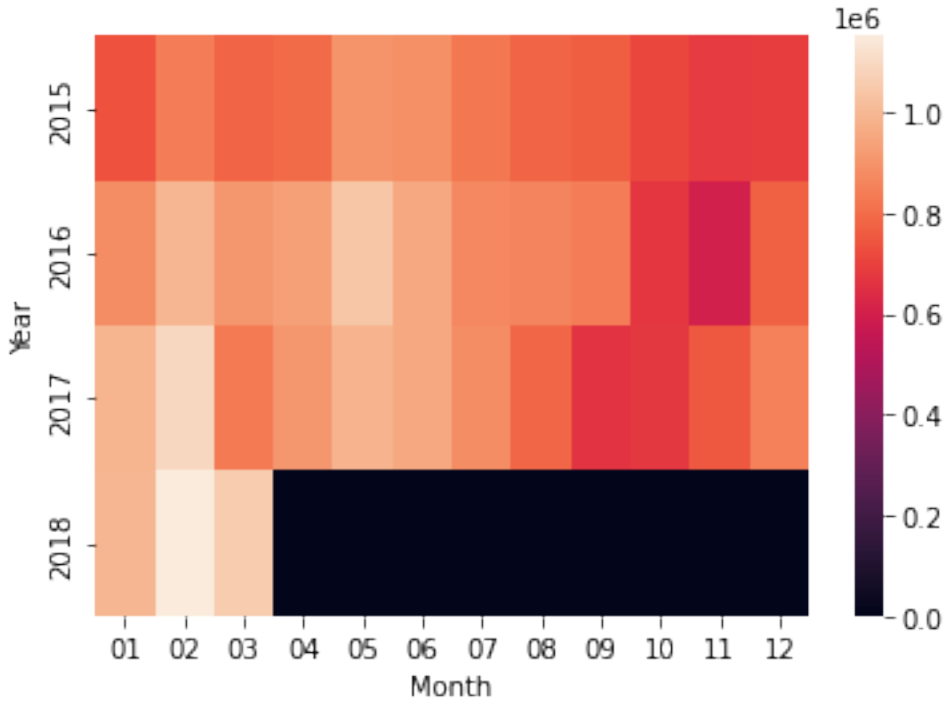
[38]:
```
ax = sns.heatmap(df.iloc[1:, [2]]);
```



[46]:
```
# %% Heatmap
# Preparación del df
df = pd.DataFrame()
df[['Year','Month','Day']] = data['Date'].str.split('-',expand=True)
df['Volume'] = data['Total Volume']
del df['Day']

df_wide = pd.pivot_table(df, index='Year', columns='Month', values='Volume')
df_wide.fillna(0, inplace=True)
df_wide

sns.heatmap(df_wide)
plt.show()
```



3. Responde las siguientes preguntas:

- **¿Hay alguna variable que no aporta información?**
  - Todas las variables aportan información, sin embargo, no todas las variables sirven directamente para la graficación porque algunas de ellas no son cuantitativas.
  - Consideramos que las variables que no aportan información precisa son la, 4046, 4225 y 4770, ya que no se conoce el contexto al que hacen referencia, haciéndolas así difíciles de usar y comprender ante un usuario/lector que no tenga conocimiento de los tipos de aguacates.

- **Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?**
  - Consideramos que si se tuviera que eliminar alguna variable sería la de año (Year), ya que se cuenta con la variable de fecha (Date) la cual aporta la misma información que la variable mencionada. También, se eliminarían las variables 4046, 4225 y 4770, ya que no aportan información útil al análisis de los datos.
- **¿Existen variables que tengan datos extraños?**
  - No, no existen variables con datos extraños ya que todos los datos que se muestran tienen una relación lógica con la propia variable.
- **Si comparas las variables, ¿todas están en rangos similares?**
  - No, porque cada variable hace referencia a información completamente independiente, a pesar de que guarden una relación entre sí, no significa que los rangos tengan que ser similares.
- **¿Crees que esto afecte el análisis de los datos?**
  - No, porque son variables independientes entre sí, es decir, los valores de una columna no afectan el de otra; a excepción de "Total Bags", la cual es la suma total de "Small Bags", "Large Bags" y "XLarge Bags".
- **¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?**
  - Las columnas "Total Bags", "Small Bags", "Large Bags" y "XLarge Bags" guardan una relación entre sí, debido a que "Total Bags" es la suma total del resto de variables antes mencionadas. Otra correlación que se puede suponer, es la que hay entre el volúmen total (Total Volume) con el total de bolsas (Total Bags), ya que a mayor volúmen habrá mayor número de bolsas.