Actividad Evaluable 3: Mapas de calor y boxplots

Equipo 3

Abiel Moisés Borja García A01654937

Aranza García Narvaéz A01654658

María Clarita Osorio Vergara A01654530

Gael Eduardo Pérez Gómez A01753336

Marco Uriel Pérez Gutiérrez A01660337

Mayo, 2022

Herramientas computacionales: el arte de la analítica

Grupo 222

Profesor:

Sergio Ruiz Loza

Instituto Tecnológico y de Estudios Superiores de Monterrey

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

```
In [2]: import matplotlib
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

        data = pd.read_csv("avocado.csv")

        data.head()
```

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

2. Describimos cada uno de los datos dentro del csv

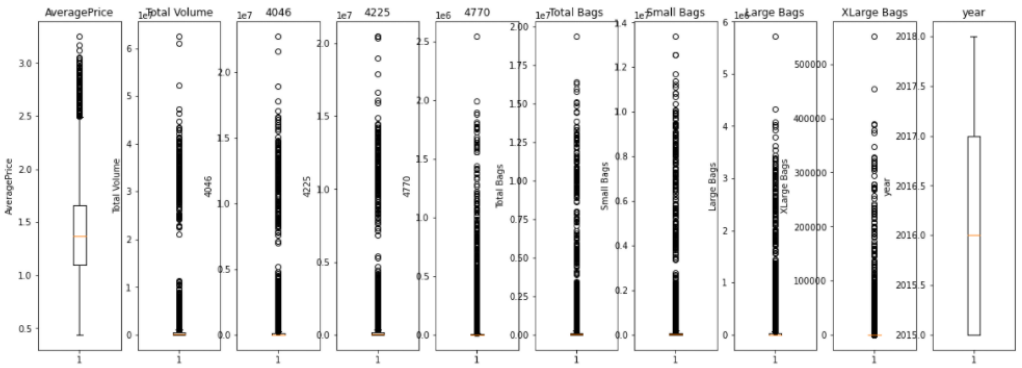
```
In [3]: data.describe()
```

Out[3]:

	Unnamed: 0	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	
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	18249.0
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	2016.1
std	15.481045	0.402677	3.453545e+06	1.264989e+06	1.204120e+06	1.074641e+05	9.862424e+05	7.461785e+05	2.439660e+05	17692.894652	0.9
min	0.000000	0.440000	8.456000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	2015.0
25%	10.000000	1.100000	1.083858e+04	8.540700e+02	3.008780e+03	0.000000e+00	5.088640e+03	2.849420e+03	1.274700e+02	0.000000	2015.0
50%	24.000000	1.370000	1.073768e+05	8.645300e+03	2.906102e+04	1.849900e+02	3.974383e+04	2.636282e+04	2.647710e+03	0.000000	2016.0
75%	38.000000	1.660000	4.329623e+05	1.110202e+05	1.502069e+05	6.243420e+03	1.107834e+05	8.333767e+04	2.202925e+04	132.500000	2017.0
max	52.000000	3.250000	6.250565e+07	2.274362e+07	2.047057e+07	2.546439e+06	1.937313e+07	1.338459e+07	5.719097e+06	551693.650000	2018.0

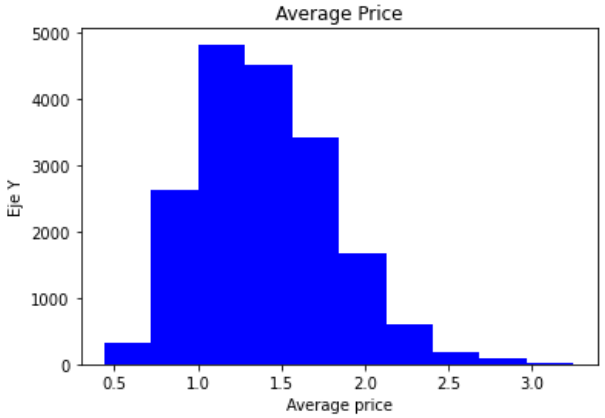
3. Para el diagramado de la caja y bigotes, definimos qué datos queremos graficar y recabar. Asimismo, definimos la longitud de las columnas, datos y las filas. Después, definimos el tamaño de las gráficas para que pueda ser legible la obtención de datos. Finalmente, mediante un for se graficaron los datos de cada una de las columnas que definimos previamente.

```
In [10]: # Cajas y bigotes
         df = data[['AveragePrice','Total Volume','4046','4225','4770','Total Bags','Small Bags','Large Bags','XLarge Bags','year']]
         ncol=len(df.columns)
         ndata=len(df)
         colgrid=ncol
         rowgrid=5

         plt.figure(figsize=(4*rowgrid,4*colgrid))
         for i,j in enumerate(df.columns):
             plt.subplot(rowgrid,colgrid,2*colgrid+i+1)
             plt.boxplot(df[j])
             plt.title(j)
             plt.ylabel(j)


         plt.show()
```

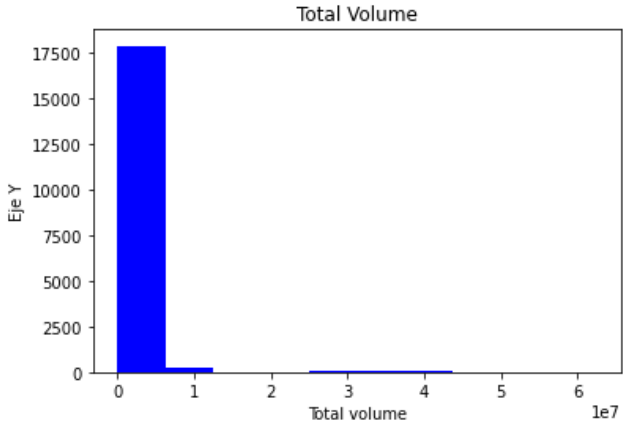


4. La graficación de histogramas nuevamente utilizamos las funciones que nos otorga la librería matplotlib. Para cada uno de los histogramas definimos los datos que queremos obtener de la data del csv y el color en el que queremos graficar. Y por otra parte, agregamos el título y el nombre de los ejes.

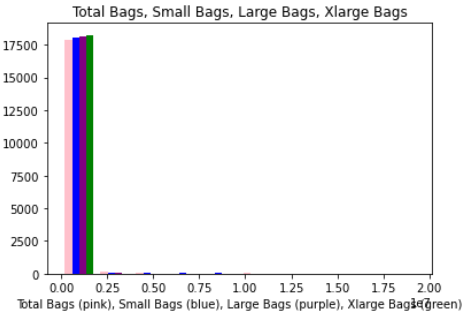
```
In [17]: # Histograma
         plt.hist(data['AveragePrice'], color='blue')
         plt.title('Average Price')
         plt.xlabel('Average price')
         plt.ylabel('Eje Y')
         plt.show()
```




```
In [18]: plt.hist(data['Total Volume'], color='blue')
         plt.title('Total Volume')
         plt.xlabel('Total volume')
         plt.ylabel('Eje Y')
         plt.show()
```



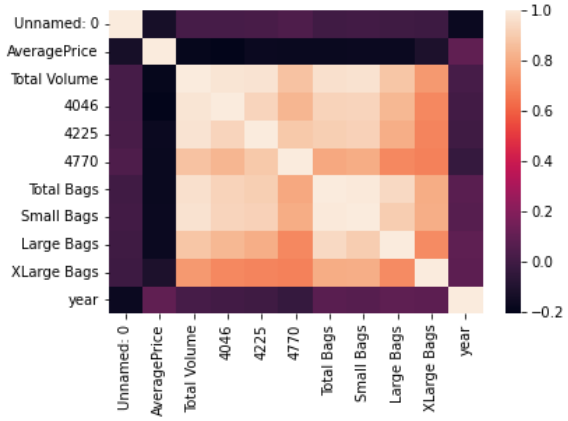
```
In [19]: bags = [data['Total Bags'], data['Small Bags'], data['Large Bags'], data['XLarge Bags']]

         label_bags = ['Total Bags','Small Bags','Large Bags','Xlarge Bags']

         colors = ['pink','blue','purple','green']

         plt.hist(bags, color = colors)

         plt.title('Total Bags, Small Bags, Large Bags, Xlarge Bags')

         plt.xlabel('Total Bags (pink), Small Bags (blue), Large Bags (purple), Xlarge Bags (green)')

         plt.show()
```



5. Para el mapa de calor utilizamos las facilidades de la librería Seaborn. Para la graficación de estos datos tuvimos en primer lugar, correlacionar los datos del csv. Una vez correlacionados, Seaborn hace todo el trabajo por nosotros al momento de graficar, nosotros como argumentos agregamos la parte de las etiquetas de los ejes.

```
In [20]: # MAPA DE CALOR
         correlacion = data.corr()
         sns.heatmap(correlacion, xticklabels = correlacion.columns, yticklabels = correlacion.columns)
```

```
Out[20]: <AxesSubplot:>
```



6. Finalmente, respondimos las preguntas de análisis

¿Hay alguna variable que no aporta información?

Cada variable aporta cierta información a su manera, puede que algunas variables sean más relevantes que otras, sin embargo, todas aportan informacion importante en la cual se puede analizar. No existe ninguna variable que no aporte información que analizar.

Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

No eliminaria ninguna variable, pues, como se mencionó todas las variables aportan información valiosa que puede ser analizada. No obstante, las variables pueden aportar menos o mucha información.

¿Existen variables que tengan datos extraños?

Dentro del archivo avocado.csv, no encontramos datos extraños o que no se puedan inferir. Sin embargo, se necesitaria más contexto del cual podemos analizar de manera certera los datos, pues, por si solos no se puede llegar a una conclusión

Si comparas las variables, ¿todas están en rangos similares?

No todas las variables se encuentran en rangos similares, pues cada una de las variables representan diferentes valores numericos de diferentes areas. Asimismo, hay variables que no tienen relación comun ni valores similares.

¿Crees que esto afecte el análisis de los datos?

La interpretación de los datos sí afecta el análisis, pues no se pueden comparar valores que no tengan relación en sí, aunque en este archivo sí cuenta con variables que tienen valores similares, no se puede analizar todos los datos entre sí, pues existe que no tengan relación.

¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

Sí, los grupos comparten ciertas caracteristicas tales como: Total Bags, Small Bags, Large Bags, XLarge Bags. Cada uno de estos grupos comparten casi las mismas caracteristicas y los datos analizados son parecidos con una diferentes cantidad de valores.

.