

Actividad Evaluable: Mapas de calor y boxplots

1. ¿Hay alguna variable que no aporta información?

Las variables 'type' y 'year'

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

Eliminaríamos 'year', ya que este se puede encontrar en la variable 'Date'.

También eliminaremos 'type' ya que proporciona muy poca información

3. ¿Existen variables que tengan datos extraños?

Sí. Por ejemplo, las variables '4046', '4225', y '4770'. No conocemos su significado, y los valores de la misma fluctúan demasiado, por lo que nos es difícil tanto tabularlas como analizarlas. Este último problema también está presente en las variables de 'Bags'

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

No. Como ya mencionamos, hay variables cuyos valores fluctúan entre unos miles y varios millones, mientras que variables como 'AveragePrice' se quedan en dos cifras

5. ¿Crees que esto afecte?

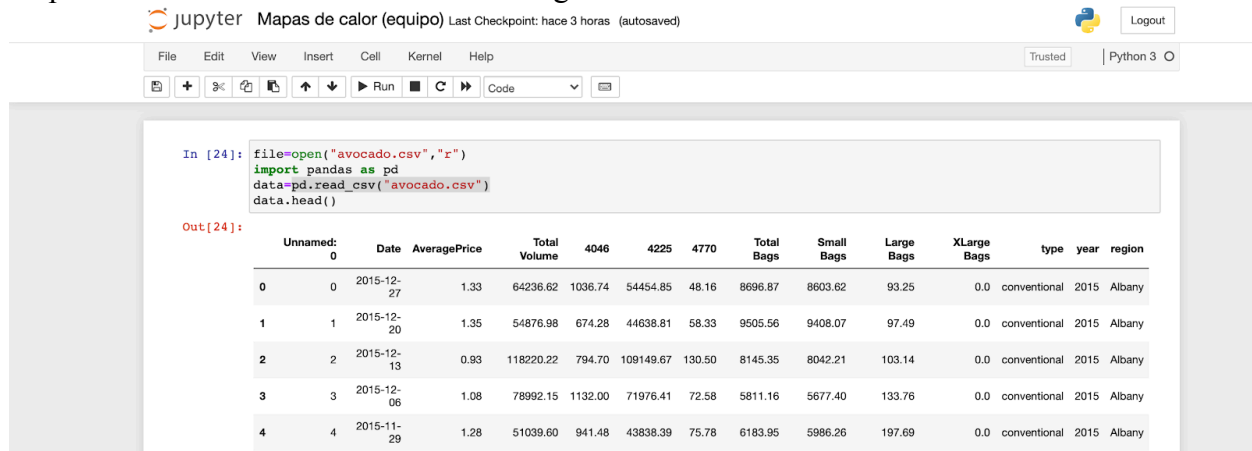
Sí, ya que a la hora de comparar datos, pueden ser confusos los análisis.

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

Los grupos de 'Bags', excluyendo 'XLarge Bags' tienen valores similares a las variables '4046' y '4225'.

Pasos para la creación de de lo pedido en la actividad .

1.Lo primero fue ver qué datos del archivo "avocado.csv", y convertir en una tabla resumiendo un poco los datos como se muestra en la siguiente tabla.

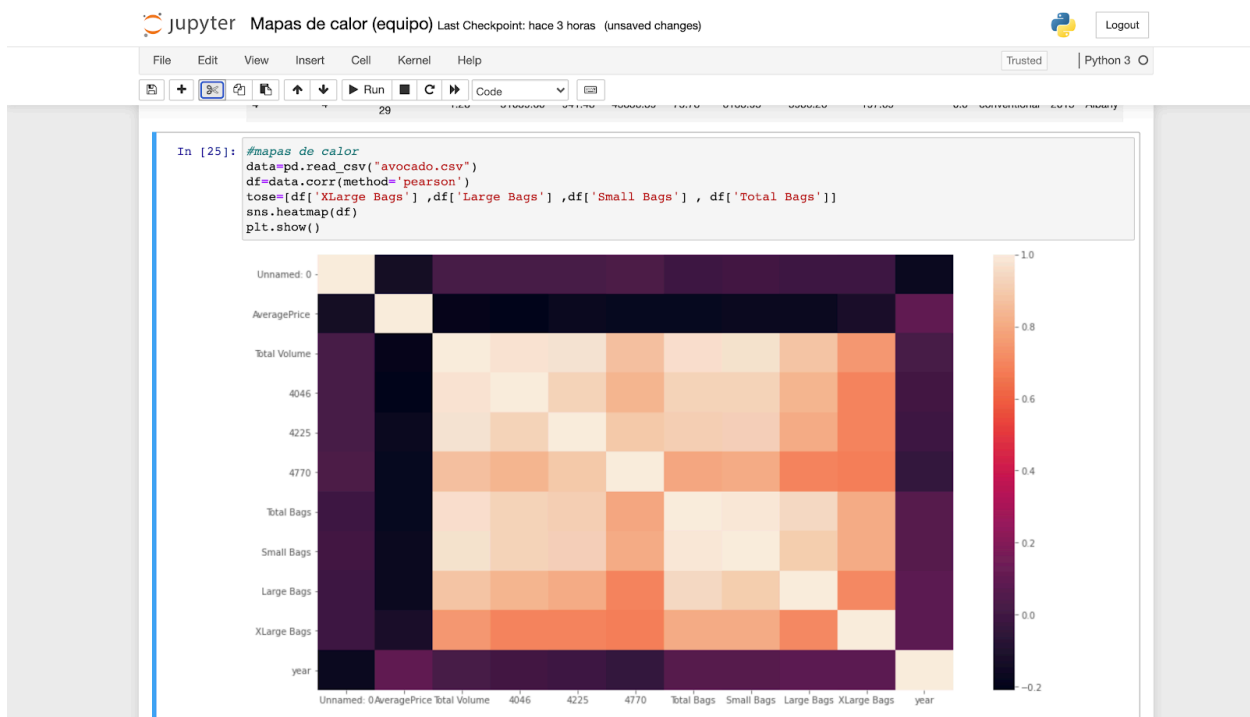


```
In [24]: file=open("avocado.csv","r")
import pandas as pd
data=pd.read_csv("avocado.csv")
data.head()
```

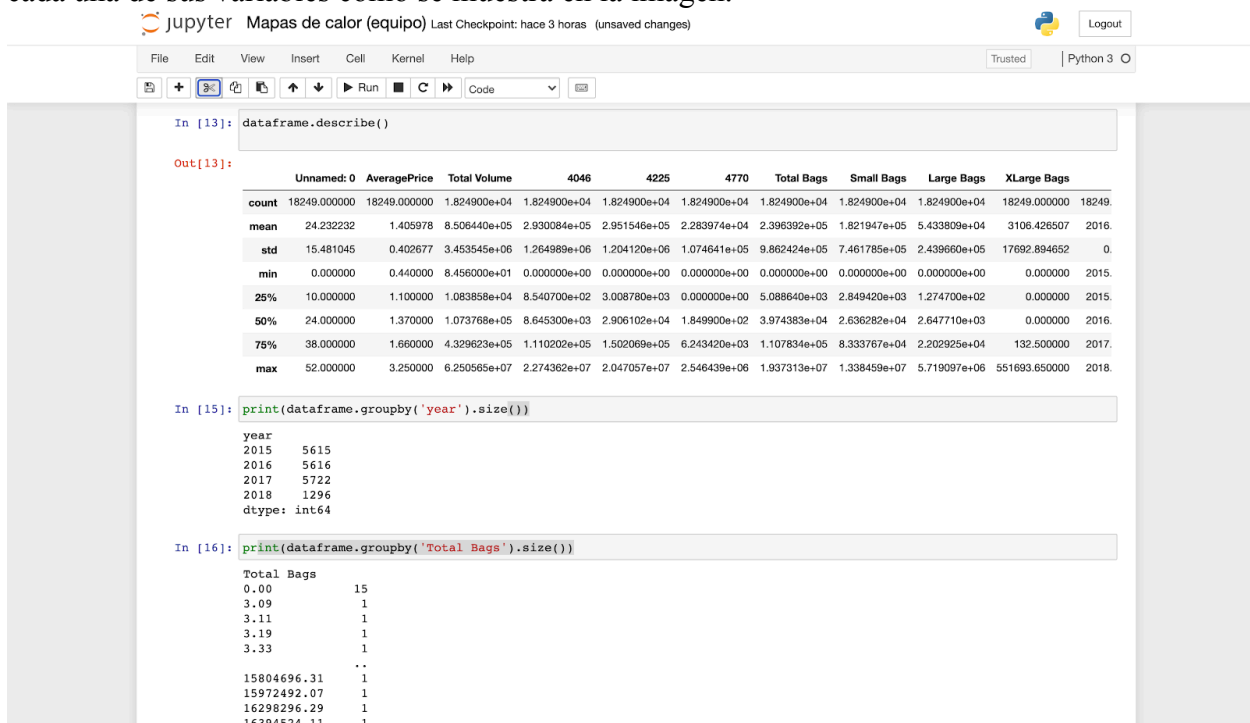
Out[24]:

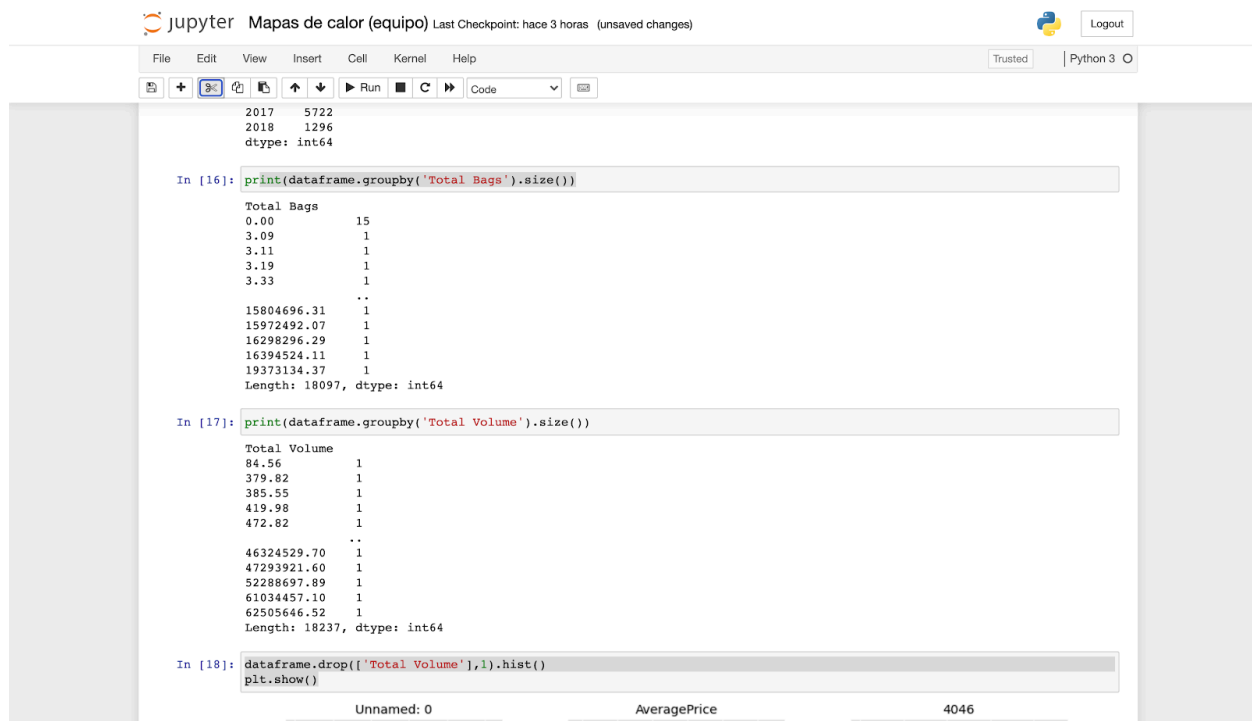
| | Unnamed: 0 | Date | AveragePrice | Total Volume | 4046 | 4225 | 4770 | Total Bags | Small Bags | Large Bags | XLarge Bags | type | year | region |
|---|------------|------------|--------------|--------------|---------|-----------|--------|------------|------------|------------|-------------|--------------|------|--------|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

2.Despues de ahí mediante funciones obtuvimos el mapa de calor con los mismos datos obtenidos en la primera tabla, en este paso se me complico un poco ya que tuve problemas con la variable data y tuve que volverla a colocar para que así fuera posible tabular.

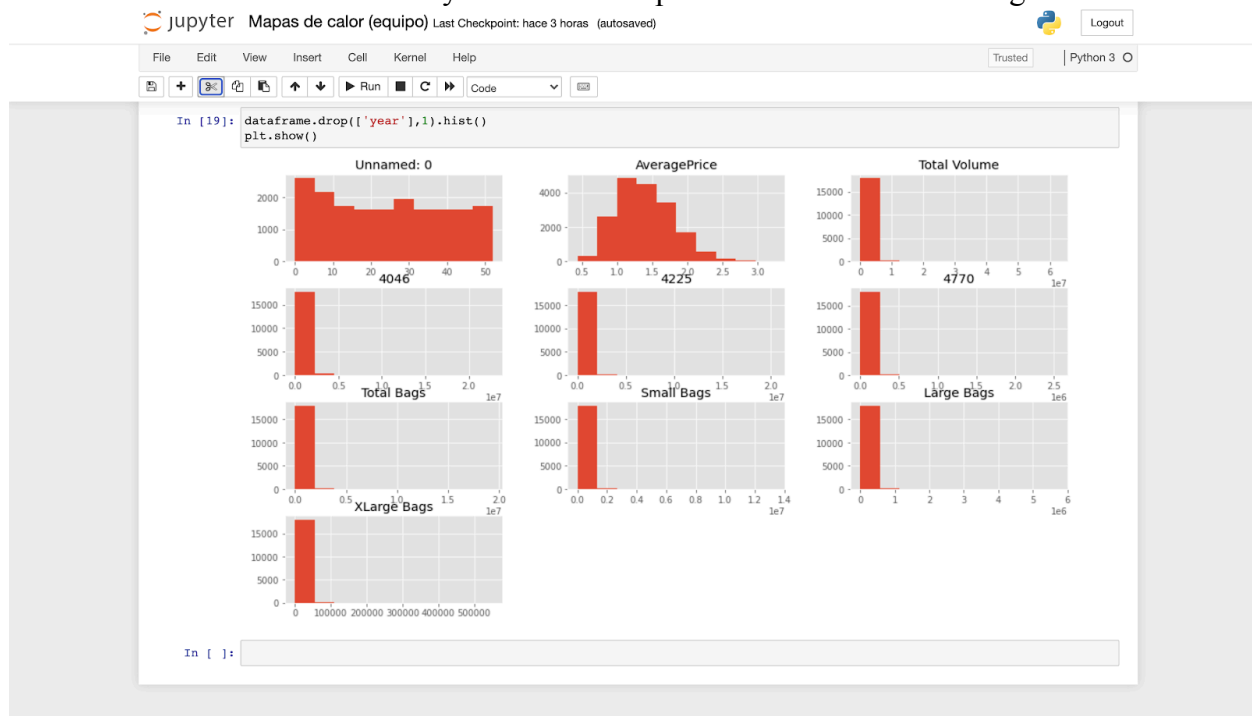


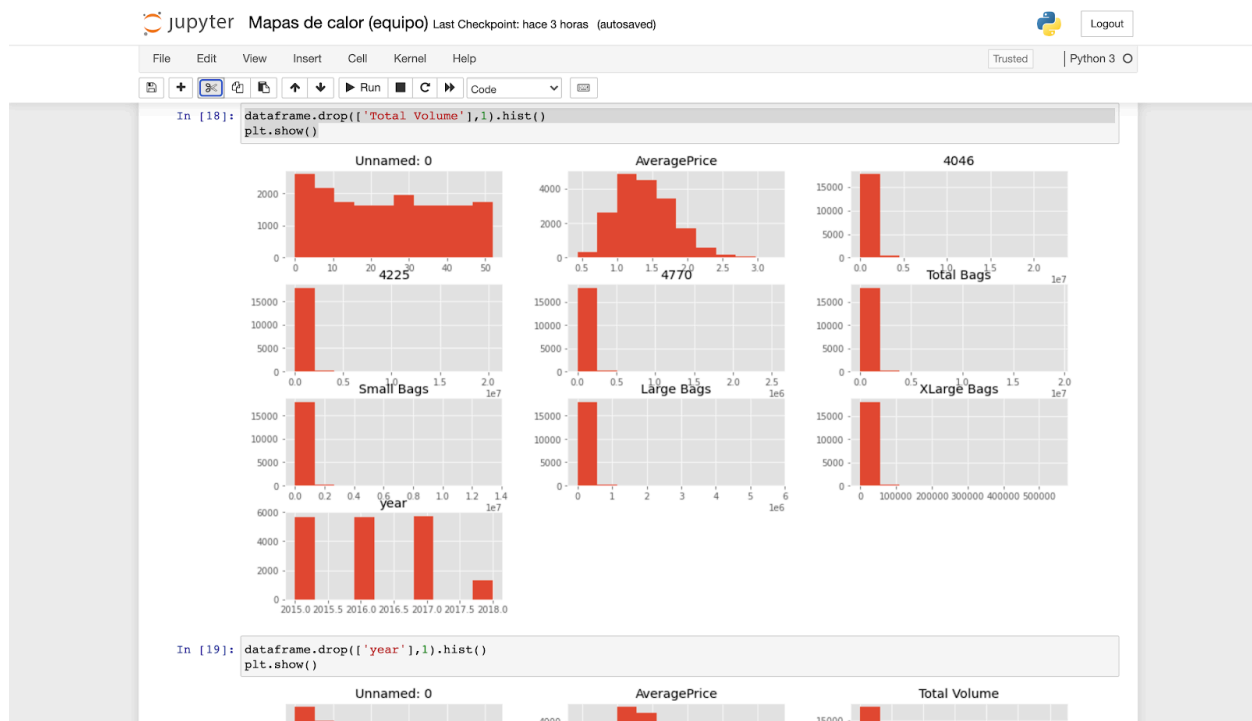
3.-Después de obtener los mapas de calor utilizamos la función de “dataframe.describe ()” para obtener de manera conjunta las variables y sus datos correspondientes y decidimos separarlos en cada una de sus variables como se muestra en la imagen.





4.- Finalmente con esas variables ya detalladas empezamos a realizar los histogramas





Para concluir con el trabajo finalmente agregamos el diagrama de cajas como se muestra en la imagen.

