Tecnológico de Monterrey

Herramientas computacionales: el arte de la analítica

TC1002S.201

Dr. Sergio Ruiz Loza

Actividad Evaluable: Mapas de calor y boxplots

Juan Ernesto Díaz Noguez	-	A01653546
Mauricio Hernández Matías		A01651328
Sara Dayana Camargo Márquez		A01652414

Ciudad de México a 28 de octubre de 2020

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

```
In [1]: import numpy as np
        import pandas as pd
        guacamoleD = pd.read_csv("avocado.csv")
        pd.set_option('display.max_rows', 10)
        print(guacamoleD)
```

```
       Unnamed: 0        Date  AveragePrice  Total Volume     4046       4225  \
0               0  2015-12-27          1.33      64236.62  1036.74   54454.85
1               1  2015-12-20          1.35      54876.98   674.28   44638.81
2               2  2015-12-13          0.93     118220.22   794.70  109149.67
3               3  2015-12-06          1.08      78992.15  1132.00   71976.41
4               4  2015-11-29          1.28      51039.60   941.48   43838.39
...           ...         ...           ...           ...      ...        ...
18244           7  2018-02-04          1.63      17074.83  2046.96    1529.20
18245           8  2018-01-28          1.71      13888.04  1191.70    3431.50
18246           9  2018-01-21          1.87      13766.76  1191.92    2452.79
18247          10  2018-01-14          1.93      16205.22  1527.63    2981.04
18248          11  2018-01-07          1.62      17489.58  2894.77    2356.13

          4770  Total Bags  Small Bags  Large Bags  XLarge Bags          type  \
0        48.16     8696.87     8603.62       93.25          0.0  conventional
1        58.33     9505.56     9408.07       97.49          0.0  conventional
2       130.50     8145.35     8042.21      103.14          0.0  conventional
3        72.58     5811.16     5677.40      133.76          0.0  conventional
4        75.78     6183.95     5986.26      197.69          0.0  conventional
...        ...         ...         ...         ...          ...           ...
18244     0.00    13498.67    13066.82      431.85          0.0       organic
18245     0.00     9264.84     8940.04      324.80          0.0       organic
18246   727.94     9394.11     9351.80       42.31          0.0       organic
18247   727.01    10969.54    10919.54       50.00          0.0       organic
18248   224.53    12014.15    11988.14       26.01          0.0       organic

       year            region
0      2015            Albany
1      2015            Albany
2      2015            Albany
3      2015            Albany
4      2015            Albany
...     ...               ...
18244  2018  WestTexNewMexico
18245  2018  WestTexNewMexico
18246  2018  WestTexNewMexico
18247  2018  WestTexNewMexico
18248  2018  WestTexNewMexico

[18249 rows x 14 columns]
```

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

   Variables: 'Unnamed: 0', 'Date', 'AveragePrice', 'Total Volume', '4046', '4225', '4770', 'Total Bags', 'Small Bags', 'Large Bags', 'XLarge Bags', 'type', 'year', 'region' .

```
In [2]: guacamoleD.columns
```

```
Out[2]: Index(['Unnamed: 0', 'Date', 'AveragePrice', 'Total Volume', '4046', '4225',
               '4770', 'Total Bags', 'Small Bags', 'Large Bags', 'XLarge Bags', 'type',
               'year', 'region'],
              dtype='object')
```

   BoxDiagram de variable ‘AveragePrice’:

```
In [189]: import matplotlib.pyplot as plt
          fig = plt.figure(figsize=(10,7))
          ax = fig.add_axes([0,0,1,1])
          bp = ax.boxplot(guacamoleD['AveragePrice'])
          plt.show()
```

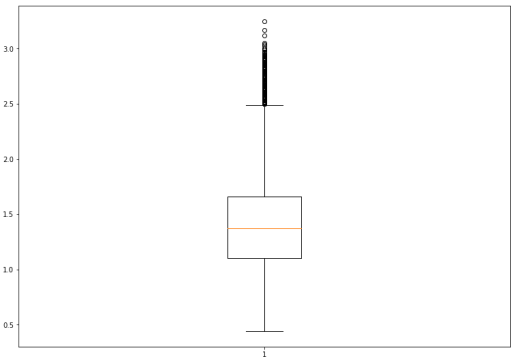

Figura 1. BoxDiagram de variable ‘AveragePrice’.

Histogramas de variables de ‘year’ y ‘AveragePrice’, respectivamente:

```
In [146]: x=guacamoleD["year"]
          plt.xlabel('year',fontsize=15)
          plt.ylabel("Frequency",fontsize=15 )
          figD=plt.hist(x, 4, width = 0.7, alpha=0.7)
          plt.show()
```

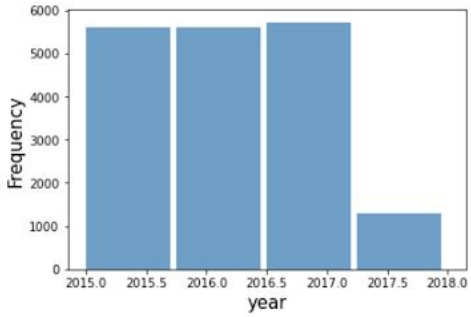

Figura 2. Histograma de variable ‘year’.

```
In [154]: x=guacamoleD["AveragePrice"]
          plt.xlabel('AveragePrice',fontsize=15)
          plt.ylabel("Frequency",fontsize=15 )
          figD=plt.hist(x, 20, width = 0.1, alpha=0.7)
          plt.show()
```

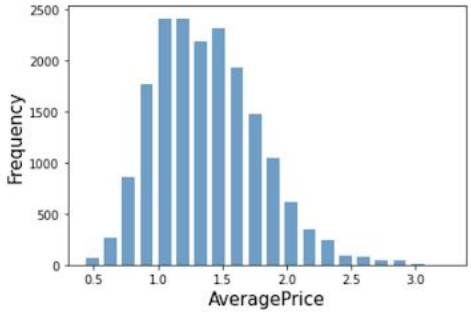

Figura 3. Histograma de variable ‘AveragePrice’.

Heatmap del número de cargamentos anuales de aguacate por ubicación. El eje *Y* representa la ubicación y el eje *X* el año:

```
In [186]: import seaborn as sns
          heatMap = pd.read_csv("avocado1.csv")
          sns.heatmap(heatMap, center=0, linewidths=0.09, linecolor="BLACK")

Out[186]: <AxesSubplot:>
```

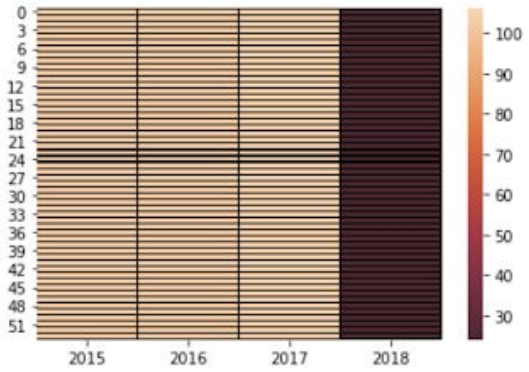

Figura 4. Heatmap del número de cargamentos anuales de aguacate por ubicación.

3. Responde las siguientes preguntas:
    1. ¿Hay alguna variable que no aporte información?

       La variables más complicadas de interpretar y que a simple vista parecen no tener una contribución de datos útiles son: ‘Unnamed: 0’, '4046', '4225' y '4770’. Sin embargo, en el caso de la variable ‘Unnamed: 0’, al analizar los datos proporcionados se encontró que la variable Unnamed representa el número de semana del año. Empezando por 0 y culminando en 51, la variable permite realizar la enumeración de la semana del año en orden ascendente, es decir, de la última semana a la primera del año.

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

       En primera instancia se eliminarían las variables 4046', '4225' y '4770’ pues no se conoce el valor que representan. Tomando en cuenta lo dicho anteriormente sobre la variable 'Unnamed: 0', otra variable que podría ser eliminada es ‘Date’, pues es complicada de analizar y es redundante al tener ‘Unnamed: 0’ y ‘year’ como variables de guía de semana y año, respectivamente.

    3. ¿Existen variables que tengan datos extraños?

       Las variables “4046”, “4225”, “4770”, al no tener contexto de los datos nuestro equipo encontró dificultad para identificar que representaban o con que variable estaban relacionadas, al igual que la variable Total Volume sabemos que la

variable nos proporciona una medida pero desconocemos que orden de magnitud se utiliza para la medida de volumen.

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

   No, pues las diferentes variables tienen órdenes de magnitud que difieren en gran medida. Además, en algunas de las variables existe una gran diferencias entre los órdenes de magnitud de los cuartiles de cada variable.

5. ¿Crees que esto afecte?

   Si, la correlación de los datos se verá afectada, así como el tipo de distribución de los datos y la desviación estándar de los datos de la mayoría de las variables de carácter cuantitativo muestra que existe una gran dispersión entre los datos de las variables, por lo que el análisis entre variables podría ser más difícil de llevarse a cabo. Esto no sería por la diferencia de magnitud entre las variables, sino por la dispersión de los datos de las variables. Podríamos encontrarnos con casos donde el coeficiente de correlación es extremadamente bajo.

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

   Sí, pues en el caso de las variables ‘Small Bags’ y ‘Large Bags’ estas realizan el mismo conteo,de diferentes productos pero es el mismo conteo de productos. Estas variables divergen en orden de magnitud (podría deberse a factores externos como la facilidad del tamaño más pequeño de venderse), sin embargo, los datos parecen ser más cercanos y parecidos al ser comparados que los de otras variables.

0.000000e+00	0.000000e+00
2.849420e+03	1.274700e+02
2.636282e+04	2.647710e+03
8.333767e+04	2.202925e+04
1.338459e+07	5.719097e+06

Figura 5. Minímo, 1er., 2do., 3er. y Máximo valor de las variables ‘Small Bags’ (columna izquierda) y ‘Large Bags (“columna derecha”).