Actividad Evaluable: Patrones con K-means

Fernando Alfonso Arana Salas A01272933

Paola Fernández Gutiérrez Zamora A01658087

Sofia Donlucas Bañuelos A01655565

Isaac Jacinto Ruiz A01658578

Santiago Gabian Perez A01658280

Grupo 222

Mayo 2022

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

Herramientas computacionales: el arte de la analítica

TC1002.S

- Carga tus datos.

```
# Importando librerias que permitiran ejecutar el algoritmo y graficar
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min

%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')

# Carga de datos
data = pd.read_csv("avocado.csv")
```

- Si determinas que alguna variable no sirve basándose en la actividad pasada, elimínala y justifica por qué quitaste o no variables.
  Las variables a eliminar que se determinaron son las correspondientes a las columnas 0, 4, 5 y 6. La variable de la columna 0, se decidió eliminar dado que no especifica a que hace referencia, no tiene un encabezado que especifique su contenido y los datos lanzados no aportan información importante para el análisis. Las variables de las columnas 4, 5 y 6 a pesar de ser de tipo float con los que se puede trabajar para su análisis estadístico no contiene un encabezado que especifique a que se refieren los datos proporcionados, por lo que a pesar de poder realizar un análisis no sabríamos que es lo que estamos analizando, si es cantidad de piezas, las bolsas no vendidas, etc. Por lo que de igual manera se decidieron eliminar.

```
# Eliminando variabes que no sirven
print(data.head())
data = data.drop(data.columns[[0, 4, 5, 6]], axis='columns')
print(data)
```

```
   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

     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

   year  region
0  2015  Albany
1  2015  Albany
2  2015  Albany
3  2015  Albany
```

```
            Date  AveragePrice  Total Volume  Total Bags  Small Bags  \
0     2015-12-27          1.33      64236.62     8696.87     8603.62
1     2015-12-20          1.35      54876.98     9505.56     9408.07
2     2015-12-13          0.93     118220.22     8145.35     8042.21
3     2015-12-06          1.08      78992.15     5811.16     5677.40
4     2015-11-29          1.28      51039.60     6183.95     5986.26
...          ...           ...           ...         ...         ...
18244 2018-02-04          1.63      17074.83    13498.67    13066.82
18245 2018-01-28          1.71      13888.04     9264.84     8940.04
18246 2018-01-21          1.87      13766.76     9394.11     9351.80
18247 2018-01-14          1.93      16205.22    10969.54    10919.54
18248 2018-01-07          1.62      17489.58    12014.15    11988.14

      Large Bags  XLarge Bags          type  year            region
0          93.25          0.0  conventional  2015            Albany
1          97.49          0.0  conventional  2015            Albany
2         103.14          0.0  conventional  2015            Albany
3         133.76          0.0  conventional  2015            Albany
4         197.69          0.0  conventional  2015            Albany
...          ...          ...           ...   ...               ...
18244     431.85          0.0       organic  2018  WestTexNewMexico
18245     324.80          0.0       organic  2018  WestTexNewMexico
18246      42.31          0.0       organic  2018  WestTexNewMexico
18247      50.00          0.0       organic  2018  WestTexNewMexico
18248      26.01          0.0       organic  2018  WestTexNewMexico
```

Las demás variables, como “date”, “type” y “region” se decidieron conservar a pesar de ser de tipo object sí aportan información importante al momento de hacer el análisis, esto debido a que al obtener las estadísticas y recurrir a estas variables complementa el hecho de saber el cuando se hizo el empaquetado del aguacate, el tipo del mismo y la región de donde salió. Esto a que pueden ser variables que influyen en los datos cuantitativos con los que se está trabajando como el total de bolsas, el tamaño y el promedio del precio. Por otro lado, variables como “AveragePrice”, “Total Volume”, “Total Bags”, “Small Bags”, “Large Bags”, “XLarge Bags” y “year” no se eliminaron dado que contienen variables numéricas de tipo float con las cuales se va a trabajar en los puntos posteriores al hacer el análisis con K-mean.

- Determina un valor de k

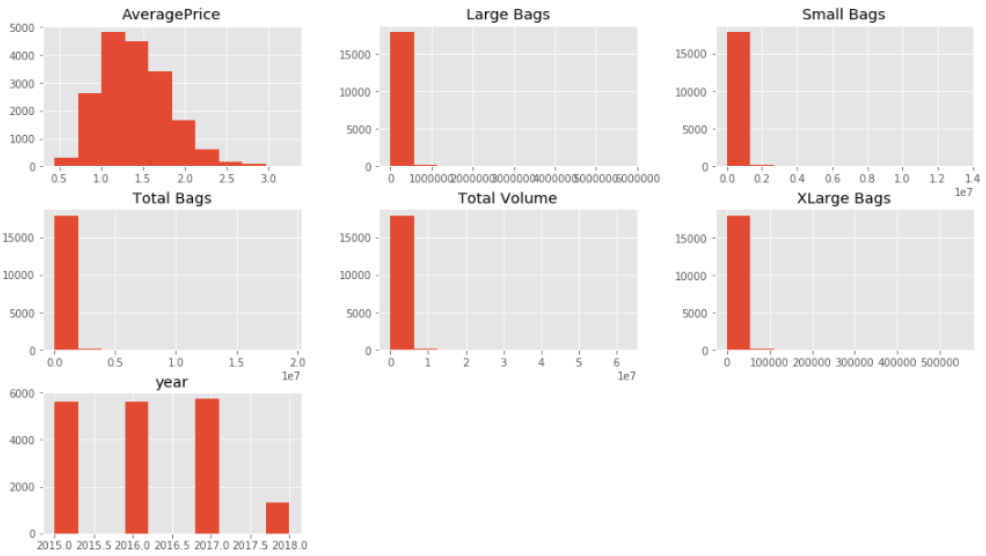
```python
# Visualizand primeros elementos
print(data.head())

# Visualizando información estadística
print(data.describe())

# Registros de total de bolsas por año
print(data.groupby('year').size())
```

```
         Date  AveragePrice  Total Volume  Total Bags  Small Bags  Large Bags  \
0  2015-12-27          1.33      64236.62     8696.87     8603.62       93.25
1  2015-12-20          1.35      54876.98     9505.56     9408.07       97.49
2  2015-12-13          0.93     118220.22     8145.35     8042.21      103.14
3  2015-12-06          1.08      78992.15     5811.16     5677.40      133.76
4  2015-11-29          1.28      51039.60     6183.95     5986.26      197.69

   XLarge Bags          type  year  region
0          0.0  conventional  2015  Albany
1          0.0  conventional  2015  Albany
2          0.0  conventional  2015  Albany
3          0.0  conventional  2015  Albany
4          0.0  conventional  2015  Albany
```

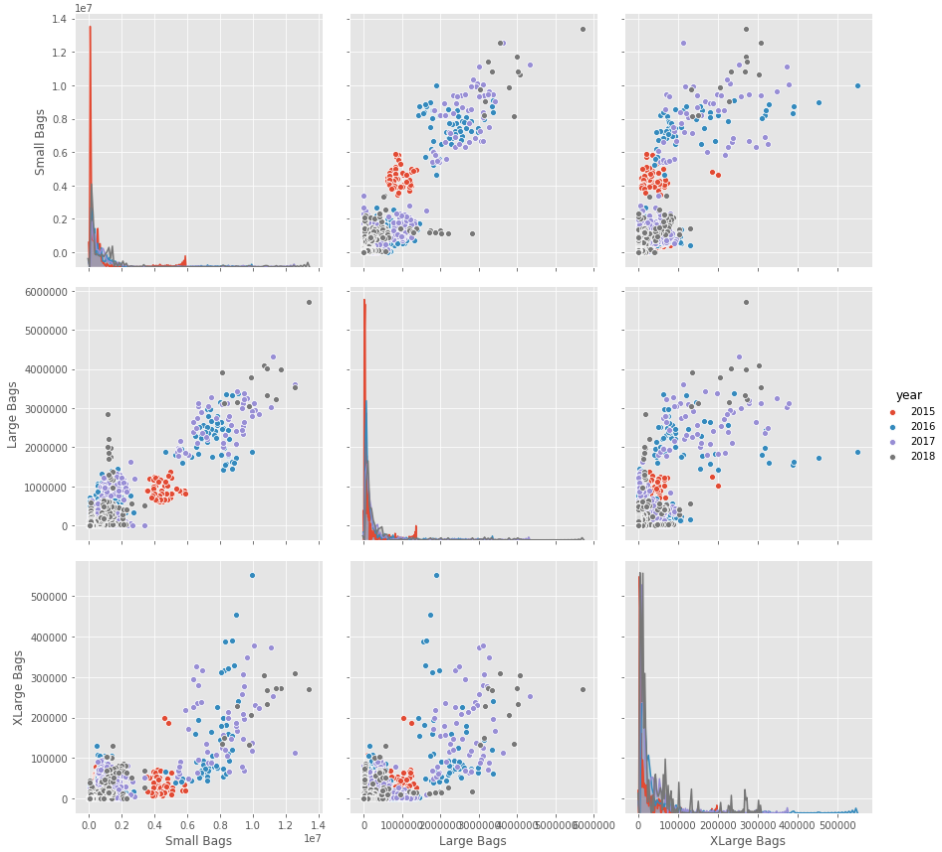
```
       AveragePrice  Total Volume    Total Bags    Small Bags    Large Bags  \
count  18249.000000  1.824900e+04  1.824900e+04  1.824900e+04  1.824900e+04
mean       1.405978  8.506440e+05  2.396392e+05  1.821947e+05  5.433809e+04
std        0.402677  3.453545e+06  9.862424e+05  7.461785e+05  2.439660e+05
min        0.440000  8.456000e+01  0.000000e+00  0.000000e+00  0.000000e+00
25%        1.100000  1.083858e+04  5.088640e+03  2.849420e+03  1.274700e+02
50%        1.370000  1.073768e+05  3.974383e+04  2.636282e+04  2.647710e+03
75%        1.660000  4.329623e+05  1.107834e+05  8.333767e+04  2.202925e+04
max        3.250000  6.250565e+07  1.937313e+07  1.338459e+07  5.719097e+06

        XLarge Bags          year
count  18249.000000  18249.000000
mean    3106.426507   2016.147899
std    17692.894652      0.939938
min        0.000000   2015.000000
25%        0.000000   2015.000000
50%        0.000000   2016.000000
75%      132.500000   2017.000000
max   551693.650000   2018.000000
```

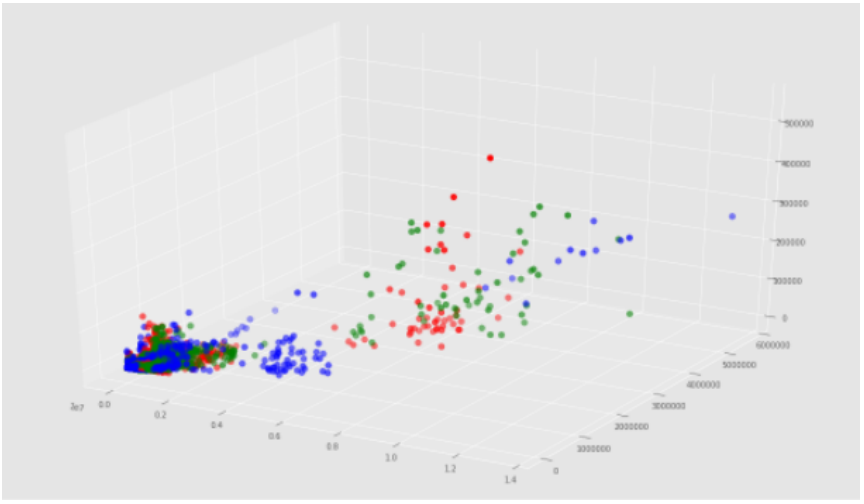
```
      year
2015  5615
2016  5616
2017  5722
2018  1296
```

```
# Visualización de datos de dispersión
data.hist()
plt.show()
```



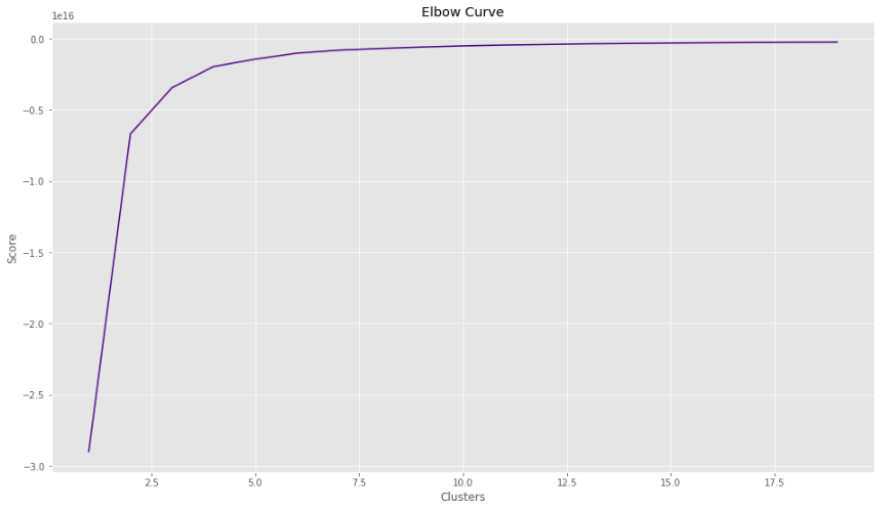

```
# Cruzando variables para obtención de pista de agrupacion con relacion a sus años
sb.pairplot(data.dropna(), hue='year',height=4,
            vars=["Small Bags","Large Bags", "XLarge Bags"],
            kind='scatter')

# Definición de entradas con variables para alimentar algoritmo
X = np.array(data[["Small Bags","Large Bags", "XLarge Bags"]])
y = np.array(data['year'])
print(X.shape)

# Gráfica 3D
fig = plt.figure()
ax = Axes3D(fig)
colores=['blue','red','green','blue','cyan','yellow','orange','black','pink','brown','purple']
asignar=[]
for row in y:
    asignar.append(colores[row-2015])
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar,s=60)
```





Valor de k=3

```
# Obteniendo el valor K
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
score
plt.plot(Nc,score, color='indigo')
plt.xlabel('Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

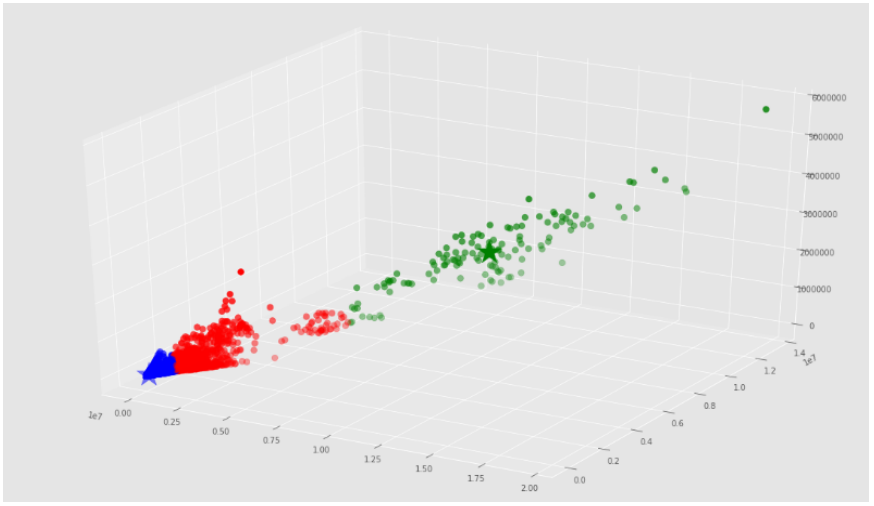



```
[[8.52980294e+04 6.45159546e+04 1.98954060e+04 8.86667923e+02]
 [1.04817264e+07 7.87924614e+06 2.44682725e+06 1.55652996e+05]
 [1.66428765e+06 1.28368037e+06 3.57724556e+05 2.28827193e+04]]
```

- Utilizando scikitlearn calcula los centros del algoritmo k-means.

```
# Ejecutando K-Means para 3 clusters
kmeans = KMeans(n_clusters=3).fit(X)

# Obteniendo etiquetas y centroids
centroids = kmeans.cluster_centers_
print(centroids)

# Grafica 3D - estrellas marcan el centro
# Prediccion de clusters
labels = kmeans.predict(X)
# Obteniendo los centros de los clusters
C = kmeans.cluster_centers_
colores=['blue','green','red',]
asignar=[]
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar,s=60)
ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c=colores, s=1000)
```

Centros de todos los clusters



Gráfica donde el algoritmo de K-means con k=3 ha agrupado a 18249 bags por su tamaño.

- ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?
  No porque estos centros hay dos que están muy cerca de otros mostrando que no son lo suficientemente diferentes. Además que el otro gran problema es que los datos que pertenecen a cada centro son muy diferentes entre sí esto mostrando que los datos de los clusters no son suficientemente parecidos entre sí y suficientemente diferentes entre clusters.

- ¿Cómo obtuviste el valor de k a usar?
  El valor de k que usamos fue tres porque ese fue el número de tipos de bolsas que había. Las grandes, las extra grandes y las chicas. Además, de que se obtuvo al hacer la gráfica “Elbow Curve” de Clusters vs Score donde se puede halla el “punto de inflexion” el cual es cercano a 3 dentro de la curva suave que se muestra. También pudimos haber evaluado otras variables así agregando más clusters y subiendo el valor de k.

- ¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?
  No sé si podrían ser más representativos o no. Tendríamos que probar con múltiples valores diferentes además que podríamos intentar usar muchos valores de k diferentes. Esto para intentar encontrar el mejor análisis posible y que podamos encontrar la mejor manera de agrupar los valores.

- ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?
  La distancia que tienen entre el cluster rojo y azul es muy pequeña. Esto es malo porque normalmente en un análisis de k-means queremos que los datos que se juntan a los centros sea muy parecida pero que los datos de diferentes centros sea muy diferente. También se puede ver como el centro verde es muy diferente a los otros dos.

- ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?
  Lo que sucedería es que los centros y toda la información se vería muy afectada y cambiaría mucho. Esto es porque una de las desventajas del k-means es que es muy vulnerable a outliers. Porque los datos se intentan juntar a centros y luego se procede

a crear nuevos centros con estos datos. Así que si hay muchos outliers esto afectaría a la creación de los centros y como todos los datos se juntan haciendo el análisis mucho más difícil.

- ¿Qué puedes decir de los datos basándose en los centros?
  Con la información obtenida del análisis, es posible concluir que se tienen datos que no son necesariamente muy parecidos entre sí dentro de cada cluster. El problema es que al tener centros muy juntos, los datos pueden quedar cerca de cualquiera de los centros que se están analizando y esto hace que las diferencias entre clusters no sea suficiente para encontrar una agrupación nueva relevante para el análisis.