



# Tecnológico de Monterrey

## Actividad Evaluable 3: Patrones con K-means

Gael González Arbesú - A01611800

Brandon Kevin Saavedra Cortes- A01748300

### ¿Qué tengo que hacer?

En esta actividad encontrarás patrones de tus datos utilizando la técnica de clustering k-means.

#### 1. Carga tus datos

```
import pandas as pd
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')

dataframe = pd.read_csv("avocado.csv")
dataframe.head()
```

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. Si determinas que alguna variable no sirve basándose en la actividad pasada, elimínala y justifica por qué quitaste o no variables.

#### 2. Si determinas que alguna variable no sirve basándose en la actividad pasada, elimínala y justifica por qué quitaste o no variables.



# Tecnológico de Monterrey

```
file Edit Selection View Go Run Terminal Help
CargaDatosPython.ipynb kmeans.ipynb Entregable_Patrones con K-means.ipynb analisisavocado.ipynb
C:\Users\pinkl\Desktop> Programas > Laboratorio > Laboratorio_A01611800 > Actividad03 > Entregable_Patrones con K-means.ipynb > Actividad Evaluable: Patrones con K-means > dataframe[["year","Unnamed: 0","XLarge Bags"]]
+ Code + Markdown ▶ Run All ⏮ Restart 🧹 Clear All Outputs 📄 Variables 📄 Outline ... Python 3.10.4

2. Si determinas que alguna variable no sirve basándose en la actividad pasada, elimínala y justifica por qué quitaste o no variables.
markdown !

dataframe[["year","Unnamed: 0","XLarge Bags"]]
[25] ✓ 0.0s Python

...
  year  Unnamed: 0  XLarge Bags
0  2015           0           0.0
1  2015           1           0.0
2  2015           2           0.0
3  2015           3           0.0
4  2015           4           0.0
...  ...           ...           ...
18244 2018           7           0.0
18245 2018           8           0.0
18246 2018           9           0.0
18247 2018          10           0.0
18248 2018          11           0.0
18249 rows x 3 columns

dataFrame=dataframe.drop(["year","Unnamed: 0","XLarge Bags"], axis=1)
dataFrame
[24] ✓ 0.0s Python

PROBLEMS 6 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER
PS C:\Users\pinkl\venv> |
```

```
file Edit Selection View Go Run Terminal Help
CargaDatosPython.ipynb kmeans.ipynb Entregable_Patrones con K-means.ipynb analisisavocado.ipynb
C:\Users\pinkl\Desktop> Programas > Laboratorio > Laboratorio_A01611800 > Actividad03 > Entregable_Patrones con K-means.ipynb > Actividad Evaluable: Patrones con K-means > dataframe[["year","Unnamed: 0","XLarge Bags"]]
+ Code + Markdown ▶ Run All ⏮ Restart 🧹 Clear All Outputs 📄 Variables 📄 Outline ... Python 3.10.4

dataFrame=dataframe.drop(["year","Unnamed: 0","XLarge Bags"], axis=1)
dataFrame
[24] ✓ 0.0s Python

...
  Date      AveragePrice  Total Volume  4046  4225  4770  Total Bags  Small Bags  Large Bags  type      region
0  2015-12-27           1.33    64236.62  1036.74  5445.485  48.16  8696.87  8603.62  93.25  conventional  Albany
1  2015-12-20           1.35    54876.98  674.28  44638.81  58.33  9505.56  9408.07  97.49  conventional  Albany
2  2015-12-13           0.93   118220.22  794.70  109149.67  130.50  8145.35  8042.21  103.14  conventional  Albany
3  2015-12-06           1.08    78992.15  1132.00  71976.41  72.58  5811.16  5677.40  133.76  conventional  Albany
4  2015-11-29           1.28    51039.60  941.48  43838.39  75.78  6183.95  5986.26  197.69  conventional  Albany
...  ...           ...           ...           ...           ...           ...           ...           ...           ...           ...
18244 2018-02-04           1.63   17074.83  2046.96  1529.20  0.00  13498.67  13066.82  431.85  organic      WestTexNewMexico
18245 2018-01-28           1.71   13888.04  1191.70  3431.50  0.00  9264.84  8940.04  324.80  organic      WestTexNewMexico
18246 2018-01-21           1.87   13766.76  1191.92  2452.79  727.94  9394.11  9351.80  42.31  organic      WestTexNewMexico
18247 2018-01-14           1.93   16205.22  1527.63  2981.04  727.01  10969.54  10919.54  50.00  organic      WestTexNewMexico
18248 2018-01-07           1.62   17489.58  2894.77  2356.13  224.53  12014.15  11988.14  26.01  organic      WestTexNewMexico
18249 rows x 11 columns

Decidimos quitar year pues tenemos date y esta es más precisa de misma forma que podemos utilizar lo planteado por year dentro de date lo cual hace que no sirva para el análisis, de misma forma quitamos Unnamed: 0 y XLarge Bags pues los datos que muestras son simples 0 para XLarge Bags y unnamed 0 no tiene relevancia alguna.
markdown

PROBLEMS 6 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER
PS C:\Users\pinkl\venv> |
```

Decidimos quitar year pues tenemos date y esta es más precisa de misma forma que podemos utilizar lo planteado por year dentro de date lo cual hace que no sirva para el análisis, de misma forma quitamos Unnamed: 0



# Tecnológico de Monterrey

y XLarge Bags pues los datos que muestras son simples 0 para XLarge Bags y unnamed 0 no tiene relevancia alguna.

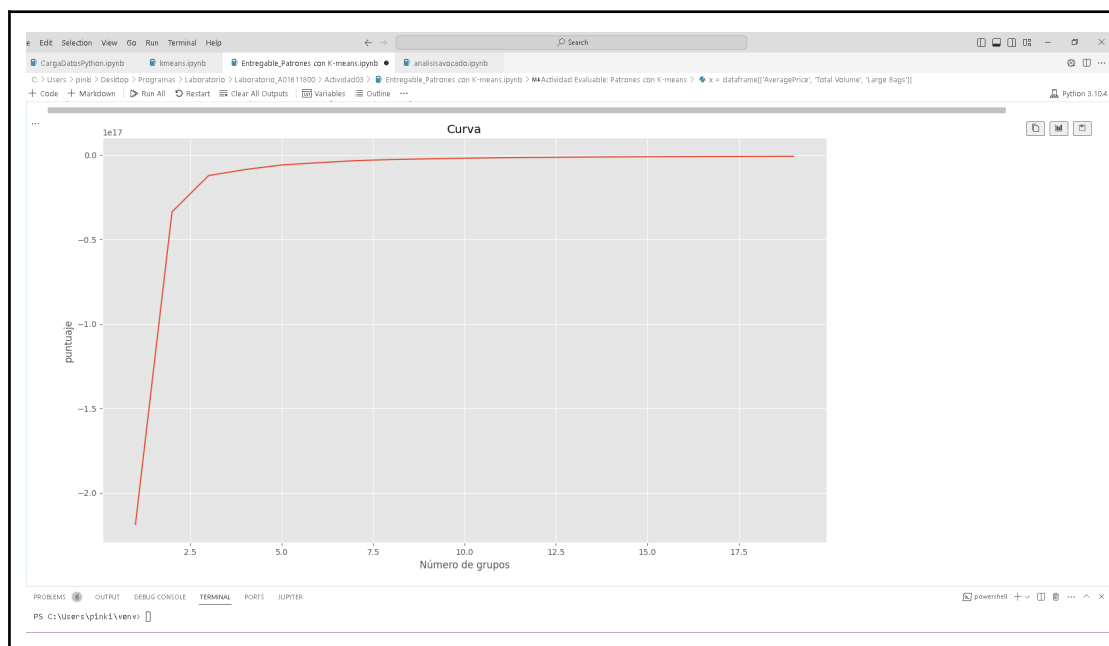
## 3. Determina un valor de k

```
File Edit Selection View Go Run Terminal Help
CargaDatosPython.ipynb kmeans.ipynb Entregable_Patrones con K-means.ipynb analisisavocado.ipynb
C:\Users\pinkie\Desktop> Programas > Laboratorio > Laboratorio_A01611800 > Actividad03 > Entregable_Patrones con K-means.ipynb > Actividad Evaluable: Patrones con K-means > x = dataframe[['AveragePrice', 'Total Volume', 'Large Bags']]
+ Code + Markdown ▶ Run All ▶ Restart Clear All Outputs Variables Outline Python 3.10.4

3. Determina un valor de k

x = dataframe[['AveragePrice', 'Total Volume', 'Large Bags']]

Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
score = [kmeans[i].fit(x).score(x) for i in range(len(kmeans))]
plt.plot(Nc, score)
plt.xlabel("Número de grupos")
plt.ylabel("puntaje")
plt.title("Curva")
plt.show()
```





# Tecnológico de Monterrey

Con esta gráfica podemos apreciar que el punto aproximado en el cual la gráfica cambia drásticamente su orientación es cerca de 2.70 en el dominio de X y sobre -0.1 en el dominio de Y, siendo más cercano a 3,0, por lo cual redondeamos tomando tal que el valor que asignaremos a K será 3.

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

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

```
kmeans = KMeans(n_clusters=3).fit(x)
centros = kmeans.cluster_centers_
print(centros)
```

```
[[1.43354656e+00 2.39568941e+05 1.77177989e+04]
 [1.09201183e+00 3.37350390e+07 2.06338732e+06]
 [1.09252680e+00 4.44383736e+06 2.64705173e+05]]
```

FutureWarning: The default value of 'n\_init' will change from 10 to 100 in version 0.25. To suppress this warning, you can explicitly set 'n\_init' to 'auto' or another number.

Basado en los centros responde las siguientes preguntas

### ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?

- Sí puesto que estos centros que sacamos nos indican ciertas tendencias dentro de los datos aunque cabe destacar que puede que no sean tan precisos debido a que los datos suelen estar muy dispersos del origen.

### ¿Cómo obtuviste el valor de k a usar?

- Utilizando comandos para Kmeans dentro de la librería asignada(ScikitLearn), de misma forma que buscamos en internet cómo implementar la misma para sacar K

### ¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?



# Tecnológico de Monterrey

- Depende del grupo de datos que se está analizando pues jugando un poco con los valores de K obtenemos gráficas que sí son más centradas al origen pero otras que no, aunque si se tuviera que dar una respuesta entre sí y no, tomaríamos el no pues realmente la aproximación no es certera y si bien cambia no lo hace significativamente.

¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?

```
Distancia entre cada centro:

# Centroides
centros = np.array([[1.43348274e+00, 2.39945956e+05, 1.77282468e+04],
                    [1.09281183e+00, 3.37359398e+07, 2.86338732e+06],
                    [1.09256332e+00, 4.44866415e+06, 2.65138828e+05]])

# Calcular la distancia entre cada par de centroides
num_centros = len(centros)
distances = np.zeros((num_centros, num_centros))
for i in range(num_centros):
    for j in range(num_centros):
        if i != j:
            distance = np.linalg.norm(centros[i] - centros[j])
            distances[i][j] = distance

# Mostrar la matriz de distancias
print("Matriz de Distancias entre Centroides:")
print(distances)

# Encontrar los centroides más cercanos entre sí
min_distance = np.min(distances[distances > 0])
closest_centros = np.where(distances == min_distance)
centro1, centro2 = closest_centros[0][0], closest_centros[1][0]
print(f"Los centros más cercanos son el Centro {centro1} y el Centro {centro2} con una distancia de {min_distance}.")

--- Matriz de Distancias entre Centroides:
[[ 0.          33557502.66870093  4215984.01617503]
 [33557502.66870093  0.          29341531.18783356]
 [ 4215984.01617503  29341531.18783356  0.          ]]
Los centros más cercanos son el Centro 0 y el Centro 2 con una distancia de 4215984.016175028.

import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np

# Datos de los centroides
centros = np.array([[1.43348274e+00, 2.39945956e+05, 1.77282468e+04],
                    [1.09281183e+00, 3.37359398e+07, 2.86338732e+06],
                    [1.09256332e+00, 4.44866415e+06, 2.64946524e+05]])
```

- Los centros más cercanos son el Centroide 0 y el Centroide 2 con una distancia de 4215984.016175028.
- Los centros 1 y 3 son los que muestran mayor cercanía a comparación del 1.

¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?

- Se encuentran más dispersos al haber mayor distancia(muchos outliers).

¿Qué puedes decir de los datos basándose en los centros?

- Que son muy similares entre sí y que los datos muestran tendencias independientemente del número de variables presentes y a gran escala los centros se encontraban cerca entre sí, concluyendo que mediante estas



# Tecnológico de Monterrey

tendencias podemos generar análisis estadísticos de los datos más relevantes y aplicables como puede ser ciertas tendencias en los precios y volúmenes.