Actividad Evaluable 4: Patrones con K-means

Equipo 1

```
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')


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

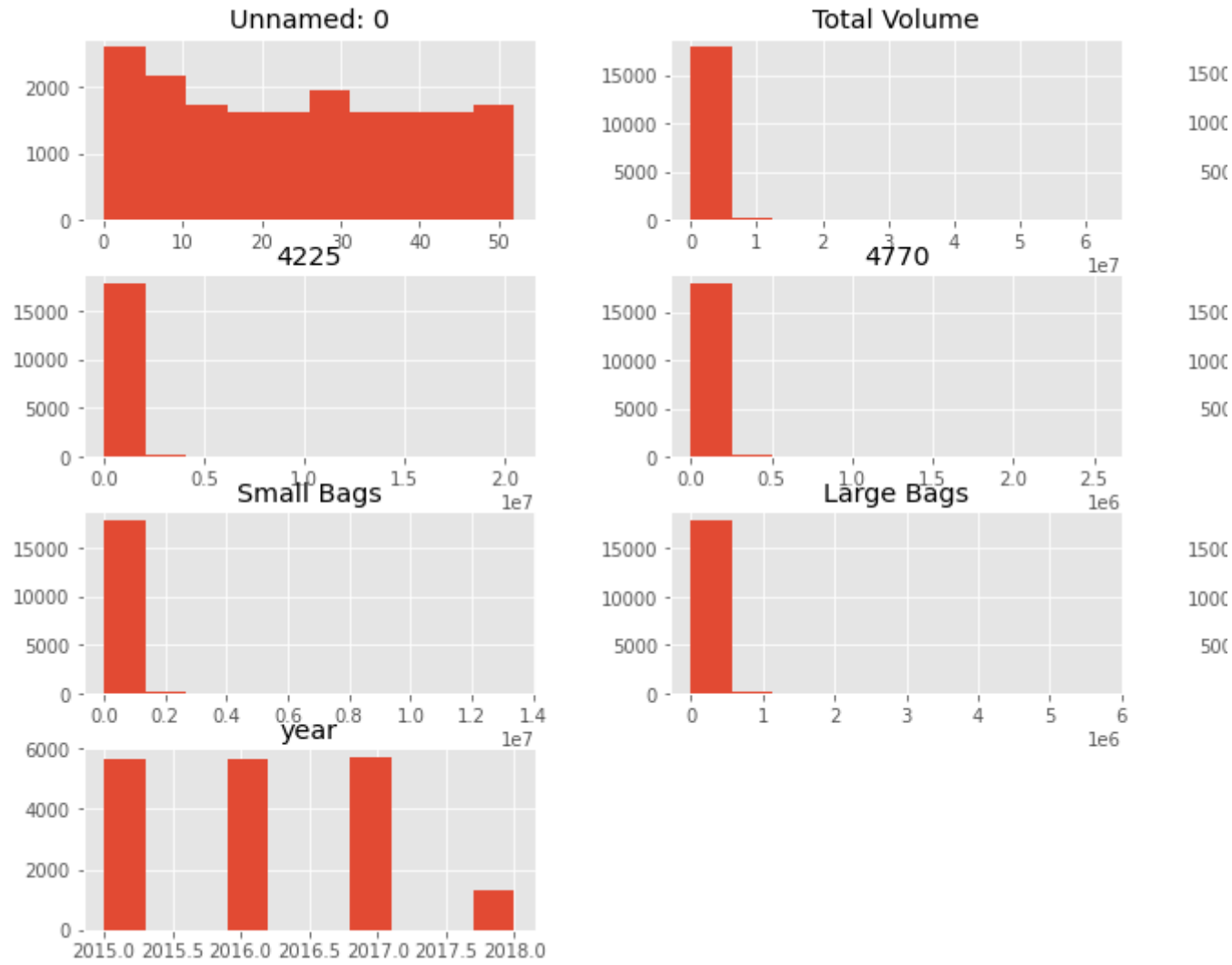
	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Tota
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	

```
dataframe.describe()
```

```
Unnamed: 0  AveragePrice  Total Volume        4046        4225        47
```

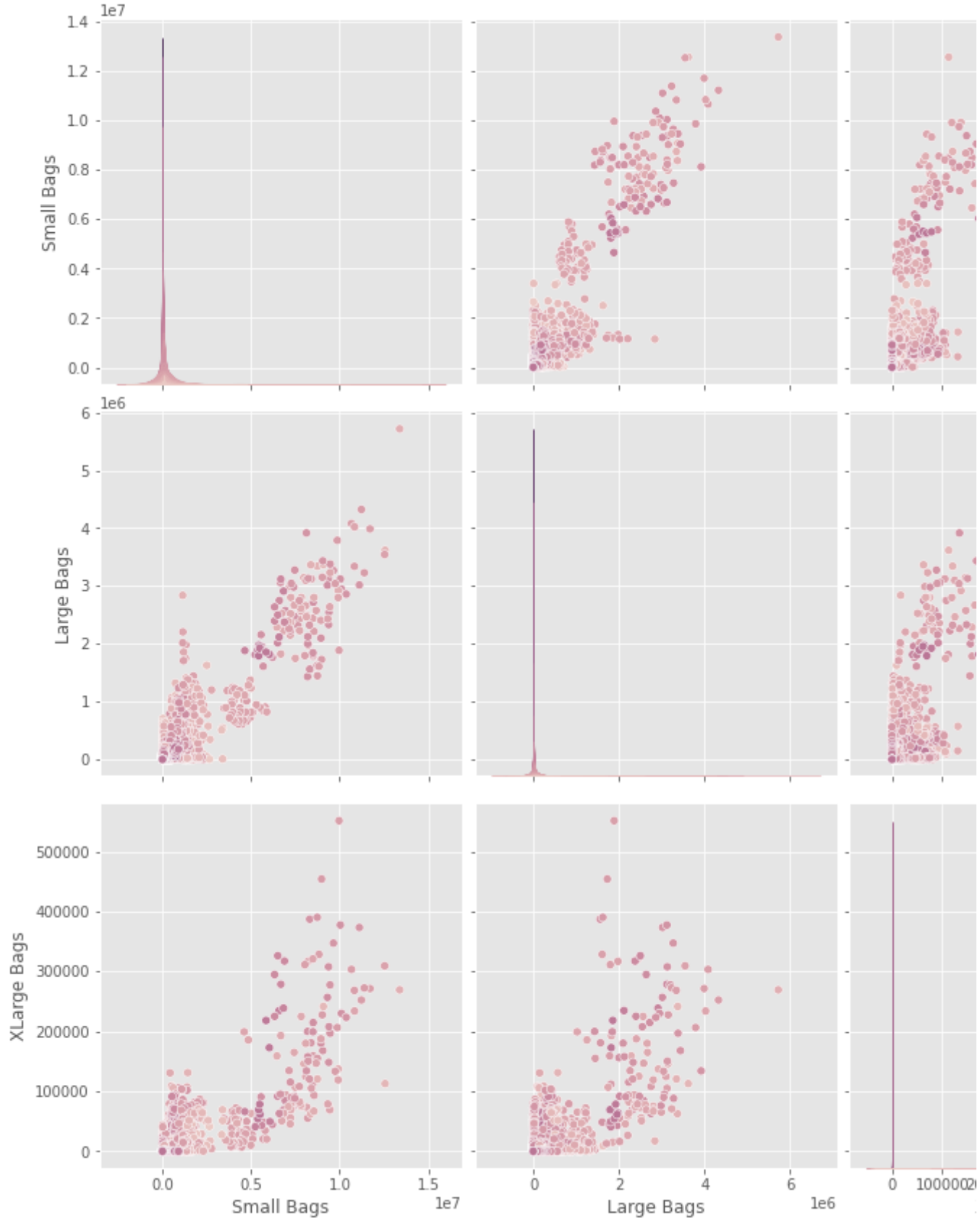
```
print(dataframe.groupby('AveragePrice').size())
dataframe.drop(['AveragePrice'],1).hist()
plt.show()
```

```
AveragePrice
0.44    1
0.46    1
0.48    1
0.49    2
0.51    5
       ..
3.04    1
3.05    1
3.12    1
3.17    1
3.25    1
Length: 259, dtype: int64
<ipython-input-86-c5dee61d50f0>:2: FutureWarning: In a future version of pandas a
  dataframe.drop(['AveragePrice'],1).hist()
```



```
sb.pairplot(dataframe.dropna(), hue='AveragePrice',size=4,vars=["Small Bags","Large Ba
```

```
/usr/local/lib/python3.9/dist-packages/seaborn/axisgrid.py:2095: UserWarning: Th
  warnings.warn(msg, UserWarning)
<seaborn.axisgrid.PairGrid at 0x7ff68801cf40>
```

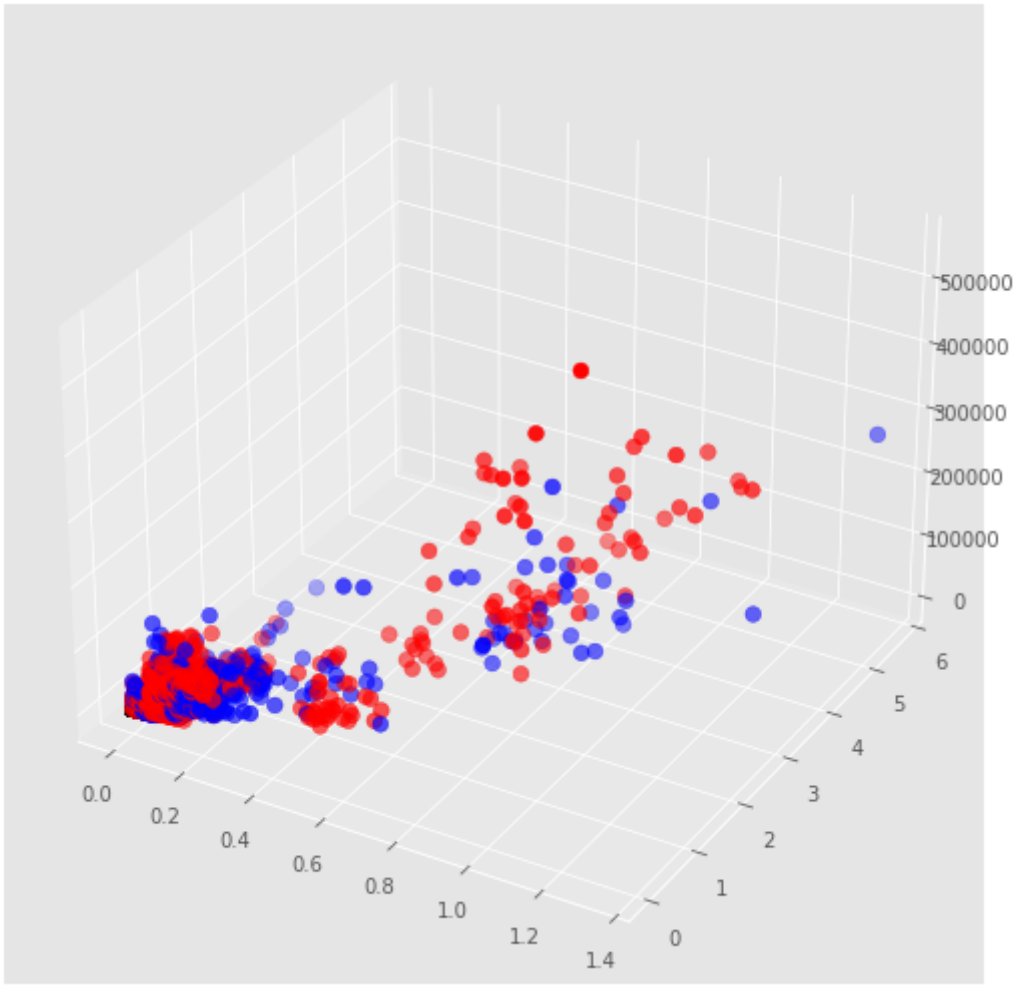


```
X = np.array(dataframe[["Small Bags","Large Bags","XLarge Bags"]])
y = np.array(dataframe['AveragePrice'])
y=y.astype(int)
print(y)
```

```
[1 1 0 ... 1 1 1]
```

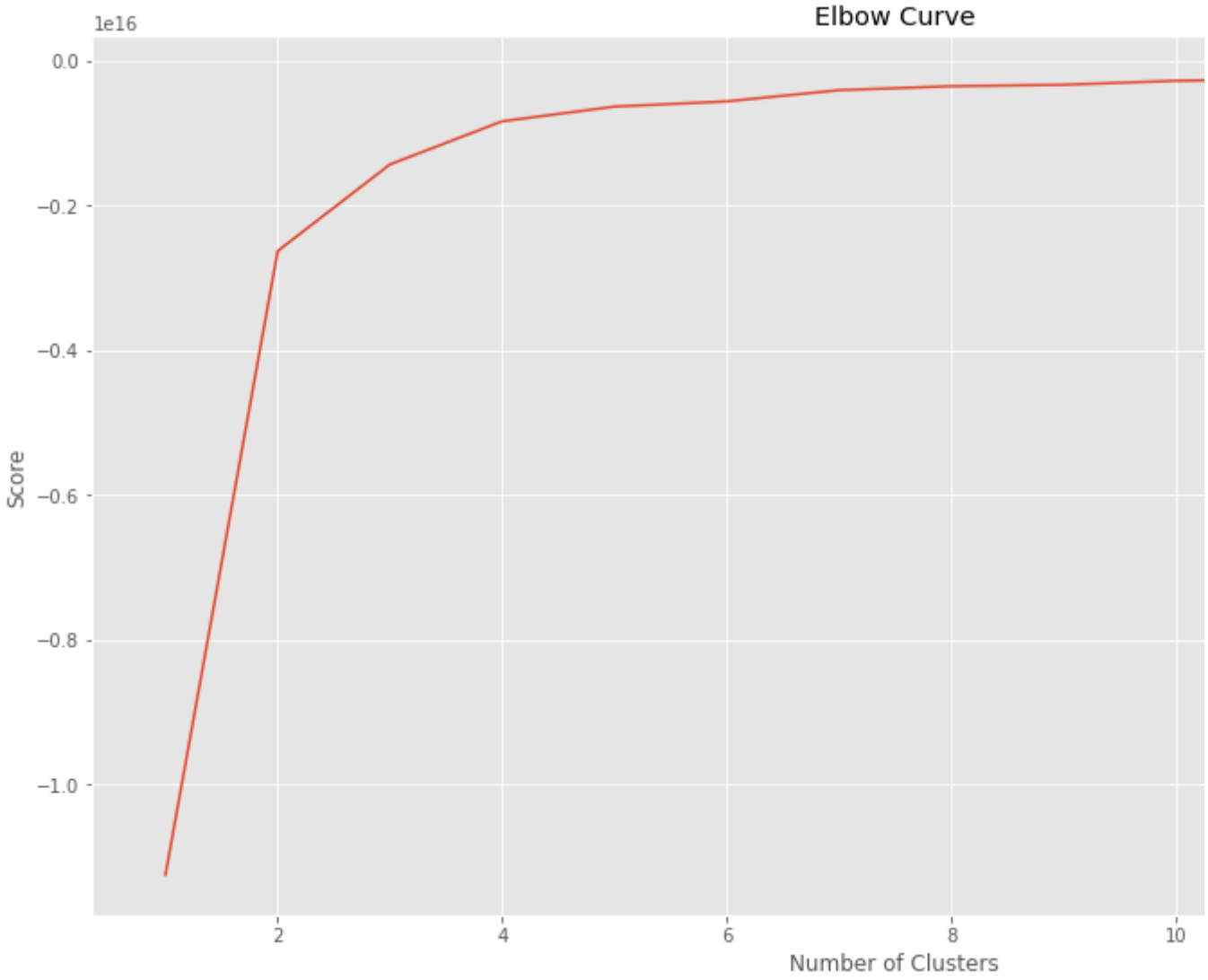
```
fig = plt.figure()
ax = fig.add_subplot(111, projection ="3d")
colores=['blue','red','green','blue','cyan','yellow','orange','black','pink','brown',
asignar=[]
for row in y:
    asignar.append(colores[row])
ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=asignar, s=60)
```

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7ff683532cd0>
```




```
Nc = range(1, 15)
kmeans = [KMeans(n_clusters=i, n_init="auto") for i in Nc]
kmeans
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
score
plt.plot(Nc,score)
```

```
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```



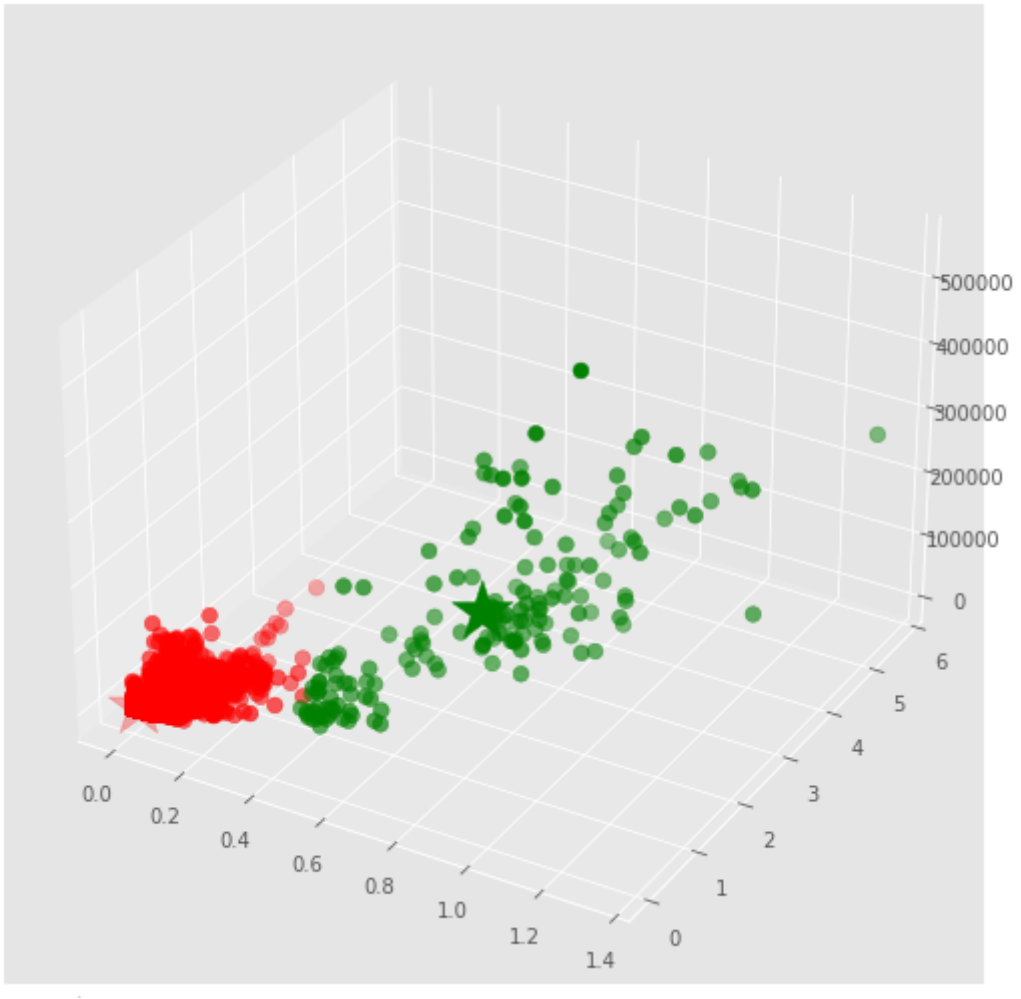

```
kmeans = KMeans(n_clusters=2).fit(X)
centroids = kmeans.cluster_centers_
print(centroids)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWar
  warnings.warn(
[[1.18842119e+05 3.56533947e+04 1.95282475e+03]
 [7.04172239e+06 2.07743166e+06 1.28013175e+05]]
```

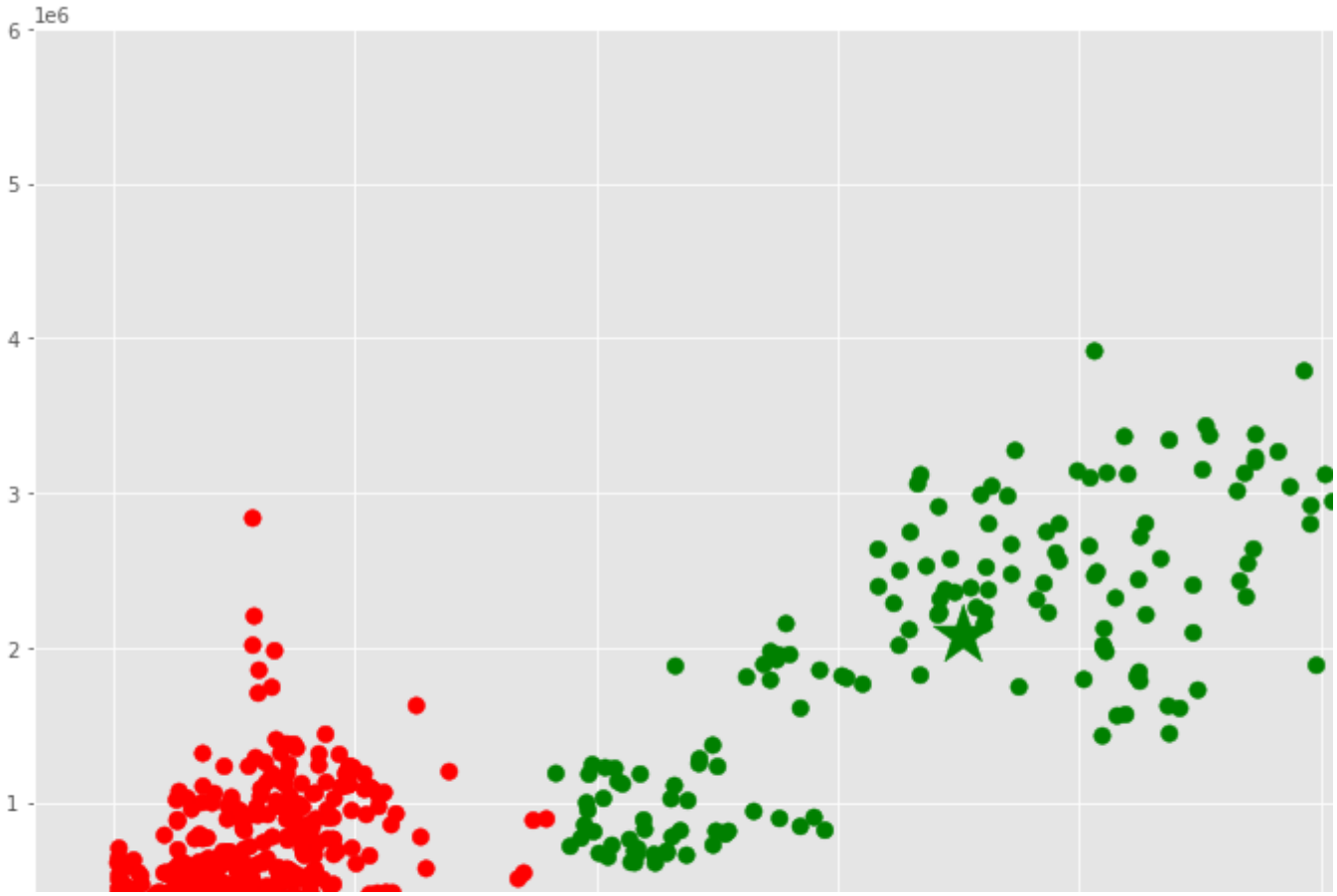
```
# Predicting the clusters
labels = kmeans.predict(X)
# Getting the cluster centers
C = kmeans.cluster_centers_
colores=['red','green']
asignar=[]
```

```
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = fig.add_subplot(111, projection ="3d")
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)
```

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7ff682c00b80>
```




```
# Getting the values and plotting it
f1 = dataframe['Small Bags'].values
f2 = dataframe['Large Bags'].values
f3 = dataframe['XLarge Bags'].values

plt.scatter(f1, f2, c=asignar, s=70)
plt.scatter(C[:, 0], C[:, 1], marker='*', c=colores, s=1000)
plt.show()
```



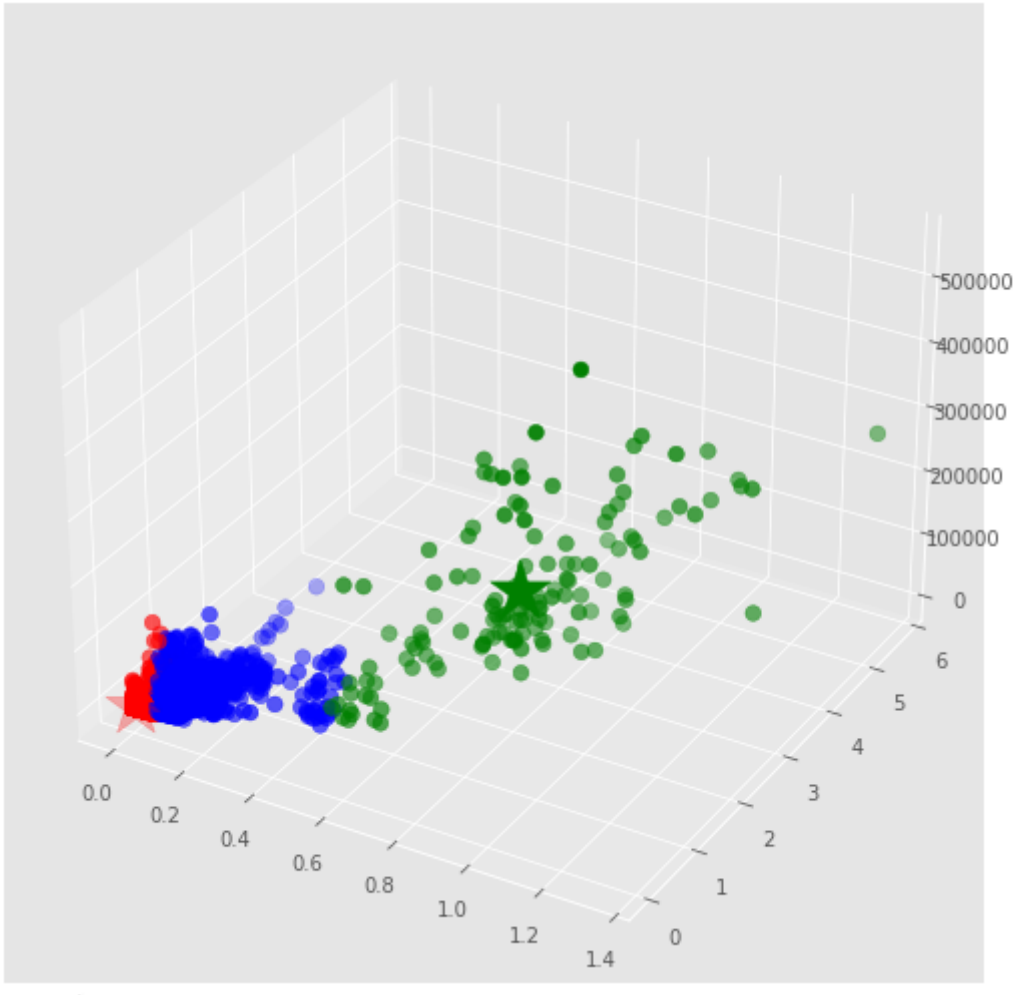
¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?

```
kmeans = KMeans(n_clusters=3).fit(X)
centroids = kmeans.cluster_centers_
print(centroids)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWar
  warnings.warn(
[[6.14708475e+04 1.99462675e+04 8.69700393e+02]
 [7.71829243e+06 2.36568410e+06 1.50320373e+05]
 [1.21702265e+06 3.27200939e+05 2.12671056e+04]]
```

```
# Predicting the clusters
labels = kmeans.predict(X)
# Getting the cluster centers
C = kmeans.cluster_centers_
colores=['red','green',"blue"]
asignar=[]
for row in labels:
    asignar.append(colores[row])

fig = plt.figure()
ax = fig.add_subplot(111, projection ="3d")
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)
```

```
<mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7fec2769ac70>
```



- José Antonio Pacheco: Concluimos que estos centros sí son representativos ya que nos ayudan a visualizar mejor la dispersión de los datos. Si usamos dos centros, existen dos grupos que se diferencian perfectamente. Por otro lado, si usamos tres centros, existen tres grupos de datos que no pueden diferenciarse de manera simple. En conlusión es más adecuado separar nuestros datos en dos grupos para facilitar su interpretación.

- Alejandro Mariacca Santin: Yo creo que los centros puedan ser representativos, los números nunca se equivocan, solo hay que saber interpretarlos. Estas interpretaciones son las que permiten afirmar o refutar teorías y/o hipótesis sobre los datos. Al mismo tiempo, nos pueden ayudar a descubrir una nueva relación en los datos que no habíamos visto.

¿Cómo obtuviste el valor de k a usar?

- Andrés Tavera: Analizamos la "elbow curve" para detectar en que puntos se pueden observar cambios rádicales. Notamos que en el punto 2 se observa el cambio más grande, pasa de un cambio constante a un cambio suave.

- José Antonio Pacheco: Para determinar nuestro valor de "k", observamos el punto donde la "Elbow Curve" presenta un cambio notorio donde la curva empieza a ser más suave. Estimamos este valor visualmente, y determinamos que es igual a 2.

- Alejandro Mariacca Santin: Para obtener el valor de K, primero tuvimos que graficar la elbow curve, que es un tipo de gráfico cuyo propósito es mostrar un punto de inflexión. Este punto de inflexión representa el punto en el cual la curva empieza a ser más suave. El valor de K es este mismo punto y puede ser calculado a simple vista. K es un valor subjetivo, cuya variación puede resultar en análisis muy distintos.

¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo? ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?

- José Antonio Pacheco: Se explicó anteriormente, los centros no serían más representativos si usaramos valores de "k" más altos. Con un valor de K=2, los centros se encuentran notoriamente separados entre sí (con una distancia de aproximadamente 0.7), lo que nos permite diferenciar los dos grupos con claridad.

- Alejandro Mariacca Santin: Si se usara un valor de K más alto, los centros serían más precisos. Por el contrario, si fuera más bajo, no serían tan confiables. Sobre su distancia, en un inicio se planteó un análisis con K=2. Se encontró que la relación es relativamente lejana. Sin embargo, cuando se usó una K=3, el grupo azul y rojo mostraron estar muy cerca y ambos están alejados del grupo verde.

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

- José Antonio Pacheco: Al presentar más outliers, tendríamos datos más dispares y difíciles de diferenciar dentro de un mismo grupo. Por este motivo, esto solamente entorpecería nuestra interpretación y visualización de los datos.

- Alejandro Mariacca Santin: Si los valores estuvieran más dispersos, los centros serían poco exactos y poco precisos. Si se graficaran con cajas y bigotes, las cajas resultarían extensas y los bigotes cortos.

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

- Duran Sanchez Pablo Ricardo

En un analisis rapido podemos obseravr que los grupos rojo y verde cuentan con muy poca similitud, ya que sus centros son muy distantes. Los datos del grupo verde, son mucho mas dispersos y su centro no se distingue tan intuitivamente. En cambio los datos del grupo rojo estan muy concentrados y su centro se puede localizar muy intuitivamente.

Andrés Tavera: Podemos observar que los grupos verde y rojo tienen una diferenca notaria, uno de estos grupos está concentrado en un área muy pequeña, mientras que la otra es muy dis

José Antonio Pacheco: Como podemos observar, los centros de los dos grupos se encuentran muy distantes entre sí. Esto nos permite diferenciar los dos grupos de datos adecuadamente.

- Alejandro Mariacca Santin: Los datos del grupo rojo y azul muestran una tendencia a estar juntos y compactados, mientras que los del grupo verde son más dispersos. A pesar de no estar graficado, hicimos una comparación entre df1 y df3, y df2 y df3. La prueba arrojó los mismos resultados que cuando K=3; es decir, grupo azul y verde fueron similares, pero no tienen nada en común con el grupo verde.

Karen Paula Mayorga Guerrero ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?

- Son representativos porque nos permiten visualizar cómo se dispersan los datos.

¿Cómo obtuviste el valor de k a usar?

- Se obtuvo gracias al método Elbow, el cual muestra la k optima a utilizar, en esta parte se decidió que el valor a usar sería 2 porque la pendiente cambia de una forma más significativa.

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

- En el caso de tener un valor alto significaría que se tendrían varios grupos similares (como se muestra en la gráfica de arriba con 3 centros en colores rojo azul y verde) por lo tanto estos grupos podrían no ser significativos. Por otro lado, si solo tuviéramos un solo centro no tendríamos una división de los datos en grupos significativos.

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

- Uno de los centros está muy cercano al valor de 0.0 mientras que el otro está en 0.7. Por lo anterior, se tendría que existe una distancia aproximada 0.69.Y, dado que solo tenemos dos centros y no más centros para comparar contra que otras distancias, entonces no se podría afirmar con certeza si se encuentran cerca o lejos.

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

- Los centros estarían sesgados por la cantidad de datos atípicos. Por lo tanto, no se tendría una correcta visualización de la dispersión de los datos.

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

- Que existe una concentración de datos en dos notables grupos, es decir que los datos se concentran en el grupo rojo o en el verde.

Double-click (or enter) to edit

Andres Tavera: Concluimos que los centros son representativos. Nos podemos dar una idea de cómo se agrupan los datos y cómo se debiden en un grupo disperso y uno dónde todos los datos se situan en un grupo pequeño. Distinguir los centros es muy complicado y dificil de hacer en el disperso, en el otro se puede intuir facilmnete. ¿Cómo obtuviste el valor de k a usar? Analizamos la "elbow curve" para detectar en que puntos se pueden observar cambios rádicales. Notamos que en el punto 2 se observa el cambio más grande, pasa de un cambio constante a un cambio suave.

¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo? ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?

Cómo mencionamos previamnte iniciamos con una k de 3 y no nos funciono, ya que no representava bien los grupos, por eso decidimos trabajar con un k de 2 porque así tenemos dos grupos: uno disperso y uno enfocado ¿Qué puedes decir de los datos basándose en los centros? Andrés Tavera: Podemos observar que los grupos verde y rojo tienen una diferenca notaria, uno de estos grupos está concentrado en un área muy pequeña, mientras que la otra es muy disperso. ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes? Entre más outlieres tengamos, más complicado es crear grupos que tengan cosas en común, es por eso que los outliers afectan. De igual manera los picos de esta gráfica se verían poco organizados y se vería un máximo y un mínimo muy exagerados.