José Antonio Mérida Castejón - 201105

Joaquín Puente - 22296

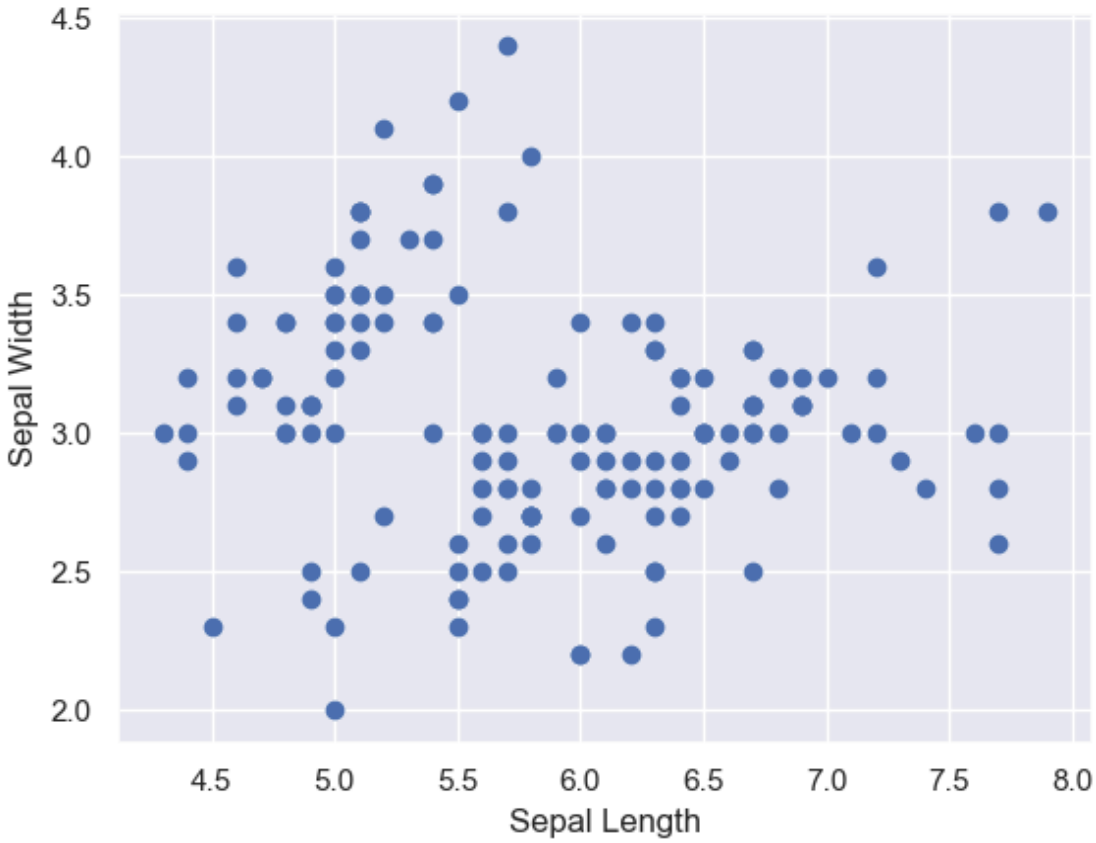
Minería de Datos

09-02-2025

Hoja de Trabajo 2. Clustering

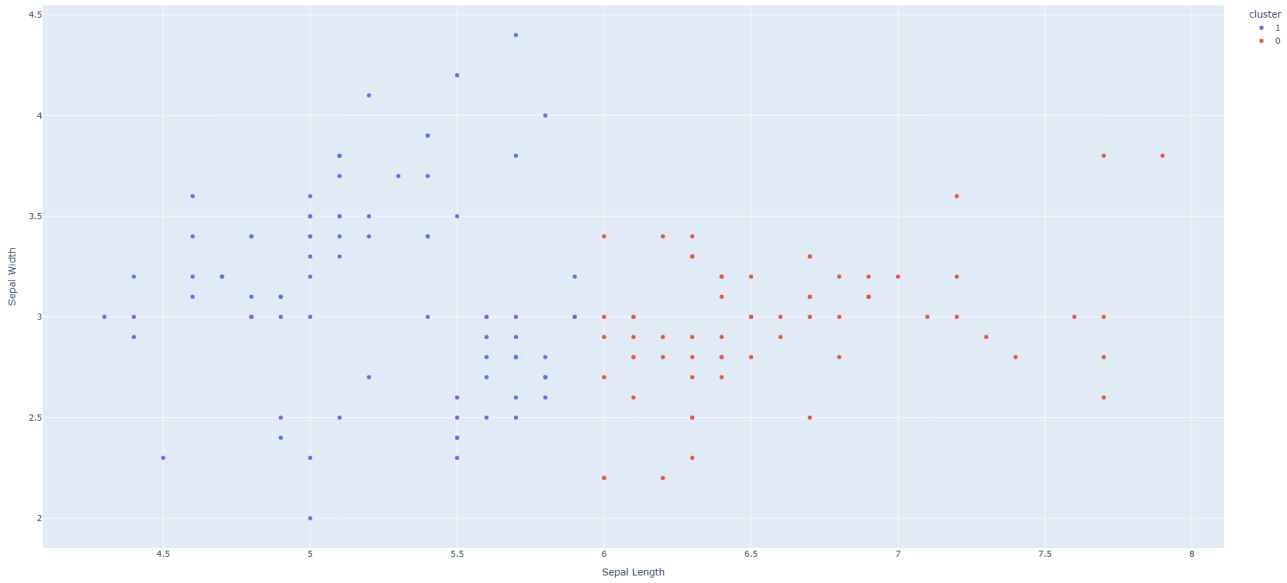
Sección 1

1. Visualicen los datos para ver si pueden detectar algunos grupos.

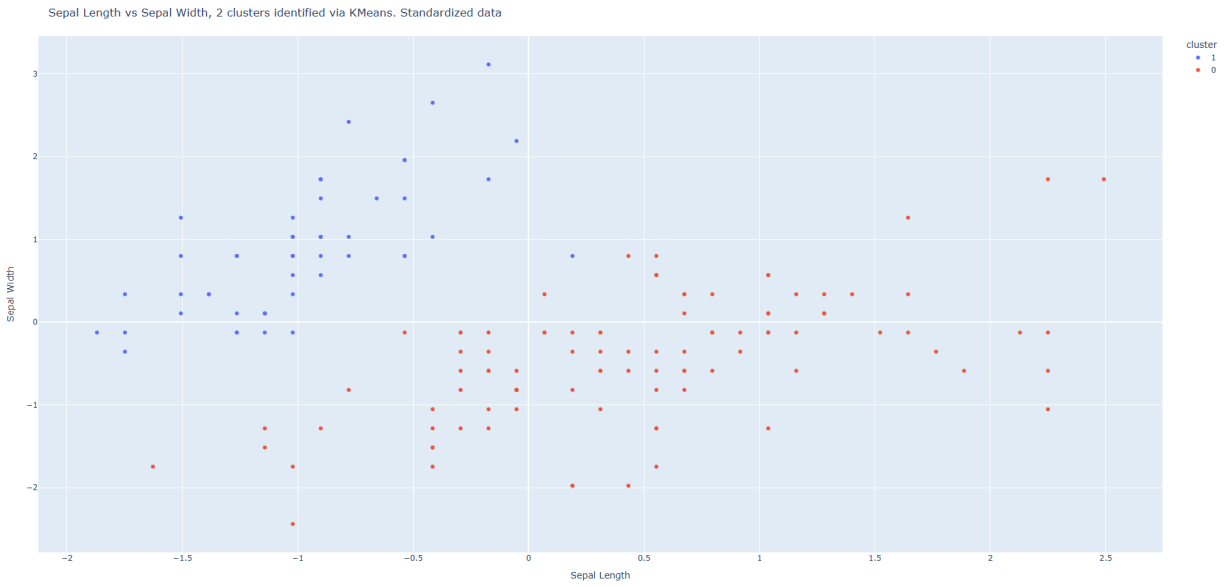


Parecen existir dos grupos marcados por una línea diagonal, los datos de la esquina superior izquierda parecen pertenecer a un cluster. En cuánto al resto de los datos, no quedan muy claras las posibles divisiones entre grupos más pequeños.

2. Creen 2 “clusters” utilizando K_Means Clustering y grafiquen los resultados.

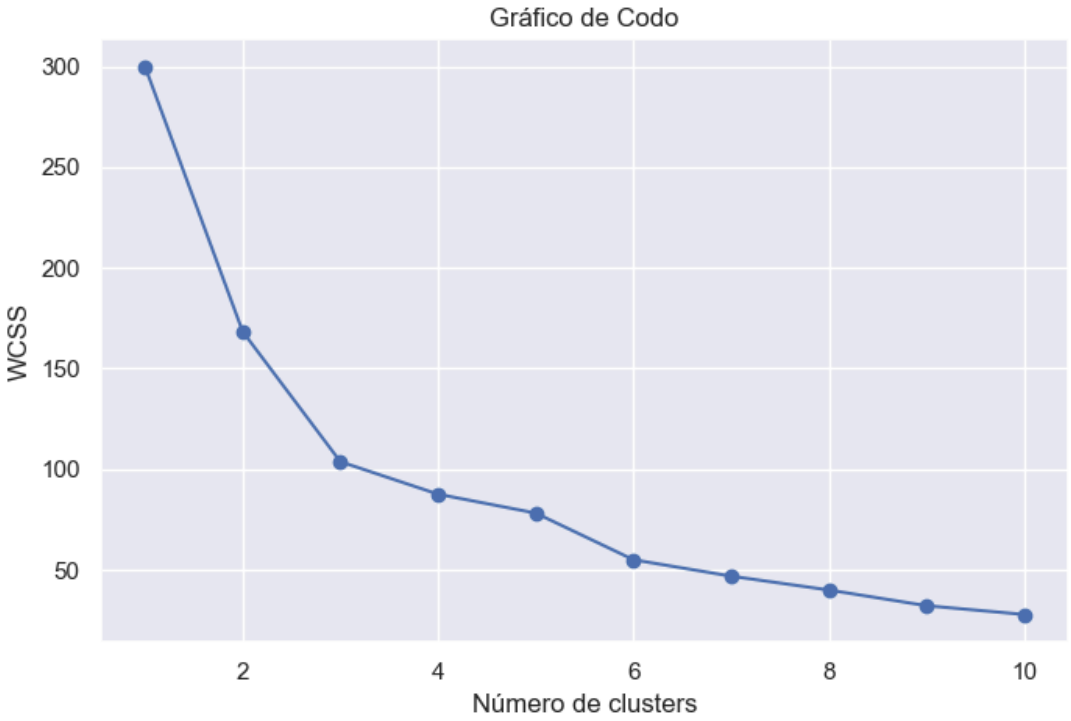


3. Estandaricen los datos e intenten el paso 2 de nuevo. ¿Qué diferencias hay, si es que hay?



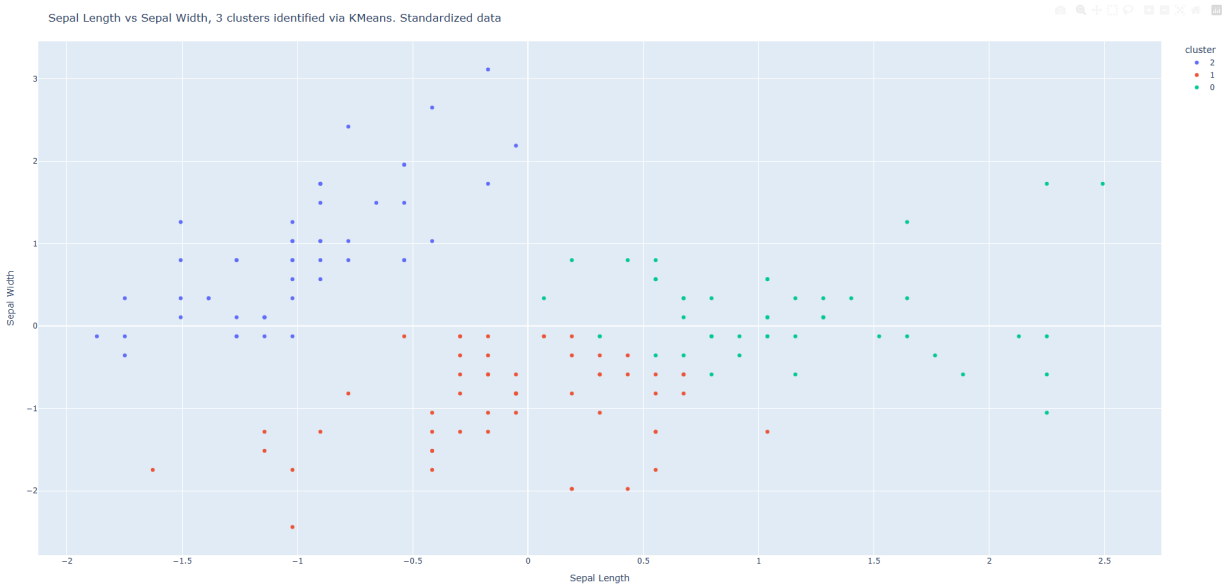
Si existe una diferencia, el cluster “0” crece significativamente. Esto se puede dar por el “peso” que se le estaba dando a la longitud del sépalo sin la estandarización de los datos. Al ser un número más grande, el algoritmo optó por una división “vertical” separando grupos en “izquierdo” y “derecho” principalmente.

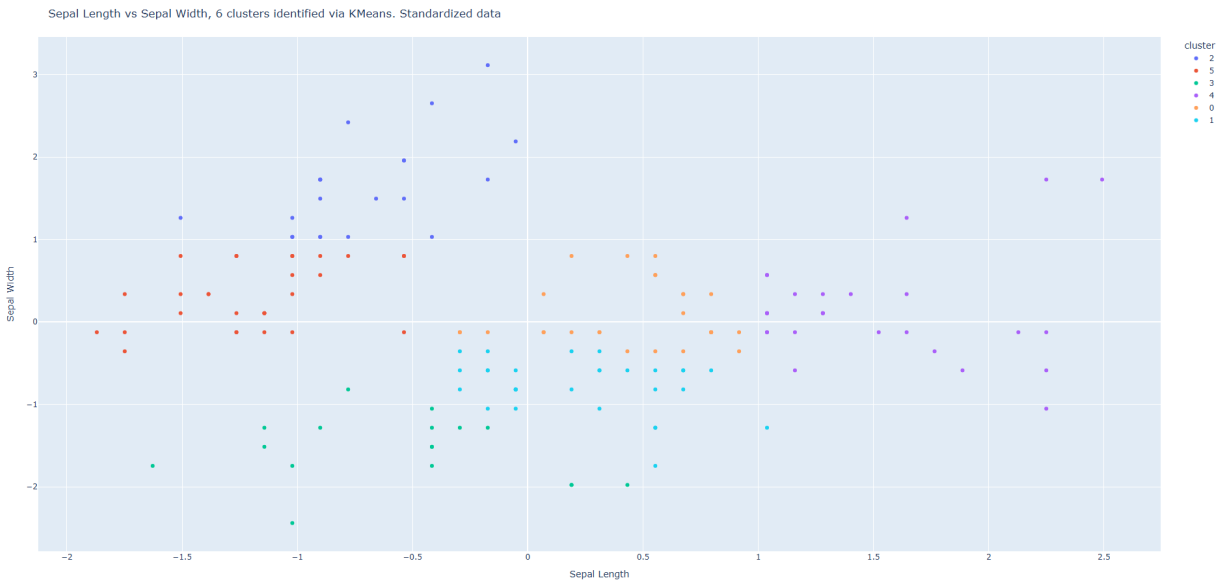
4. Utilicen el método del “codo” para determinar cuantos “clusters” es el ideal (Rango de 1 a 10)



El número ideal de clusters es de 3, ya que el descenso de WCSS se empieza a aplanar en este punto.

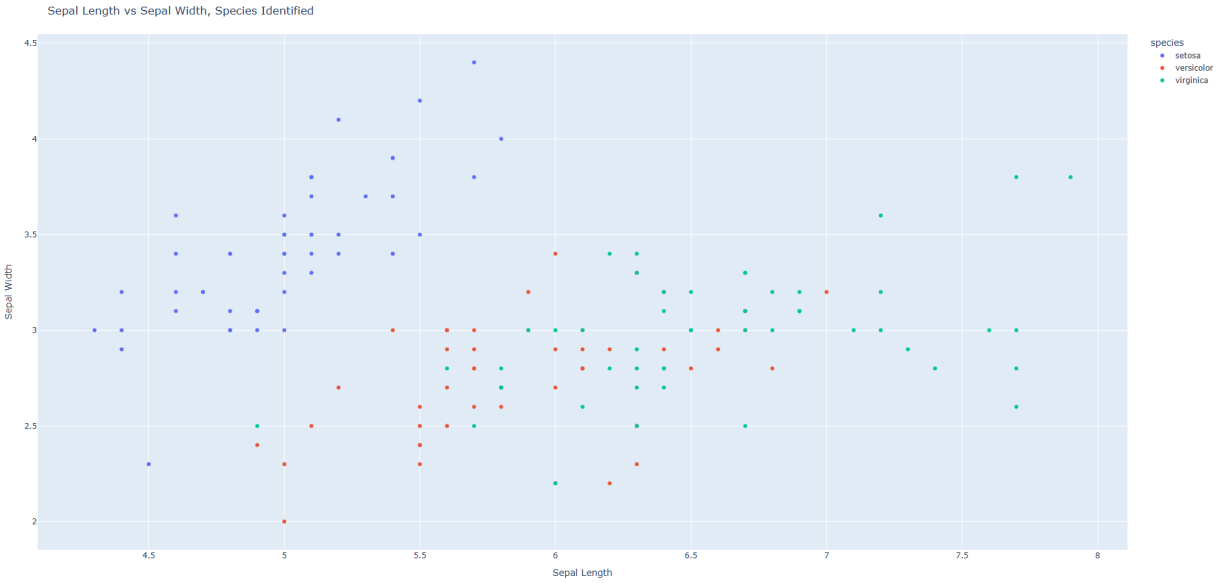
5. Basado en la gráfica del “codo” realicen varias gráficas con el número de clusters (unos 3 o 4 diferentes) que creen mejor se ajusten a los datos





Elegimos graficar con 3, 4 y 6 clusters. Esto debido a que la gráfica no tiene puntos de inflexión claramente definidos más hacia la derecha.

6. Comparen sus soluciones con los datos reales (Respondido en el siguiente inciso)



7. ¿Funcionó el clustering con la forma del sépalo?

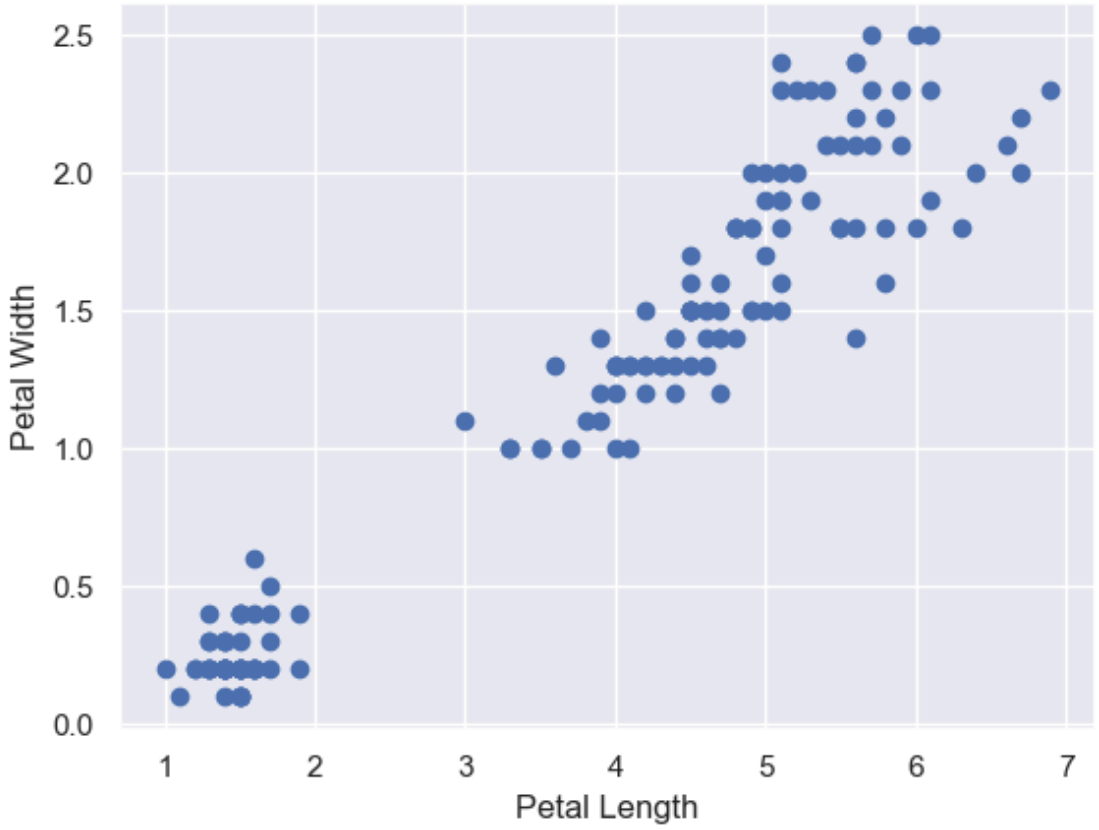
Observando las gráficas, la especie setosa parece haber sido correctamente agrupada dentro de un cluster. Esto se debe a las diferencias marcadas que tiene en comparación a las demás. Sin embargo, las especies virginica y versicolor parecen “mezclarse” bastante entre ellas

dependiendo del individuo. Esto dificulta el que funcione el algoritmo de clustering, ya que este agrupa individuos similares y separa a los disimilares. En este caso por ejemplo, podemos ver puntos rojos (individuos versicolor) cerca del límite inferior y el límite superior en cuanto a longitud del sépalo de la población total. Este análisis se respalda con un ARI de 0.53, esto nos indica que el clustering logró identificar ciertos datos pero se queda corto en cuánto a precisión cercana a la realidad. Se encuentra en el punto medio de una agrupación aleatoria y los datos 100% precisos. En cuánto a las gráficas con un mayor número de clusters, estas sufren de la misma limitación que la de 3 clusters. Las líneas "borrosas" entre especies hacen muy difícil agruparlas, y a pesar de tener un número mayor de clusters se siguen observando agrupaciones que no van acorde a los datos reales.

En conclusión, el clustering con la forma del sépalo no funcionó con la precisión que se esperaba.

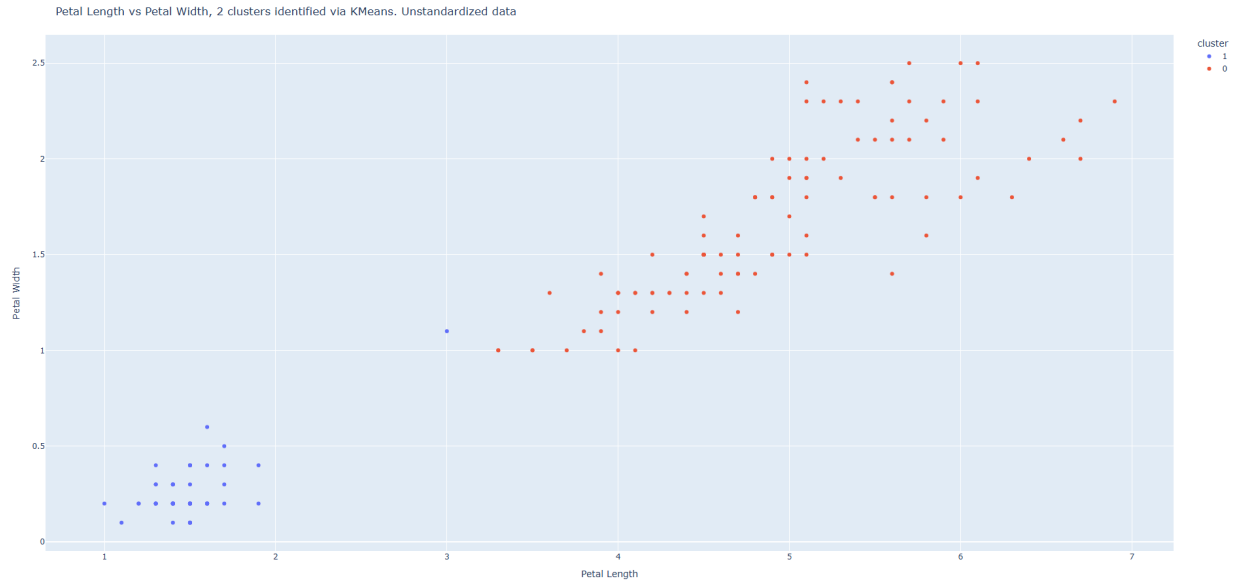
Sección 2

1. Visualicen los datos para ver si pueden detectar algunos grupos.

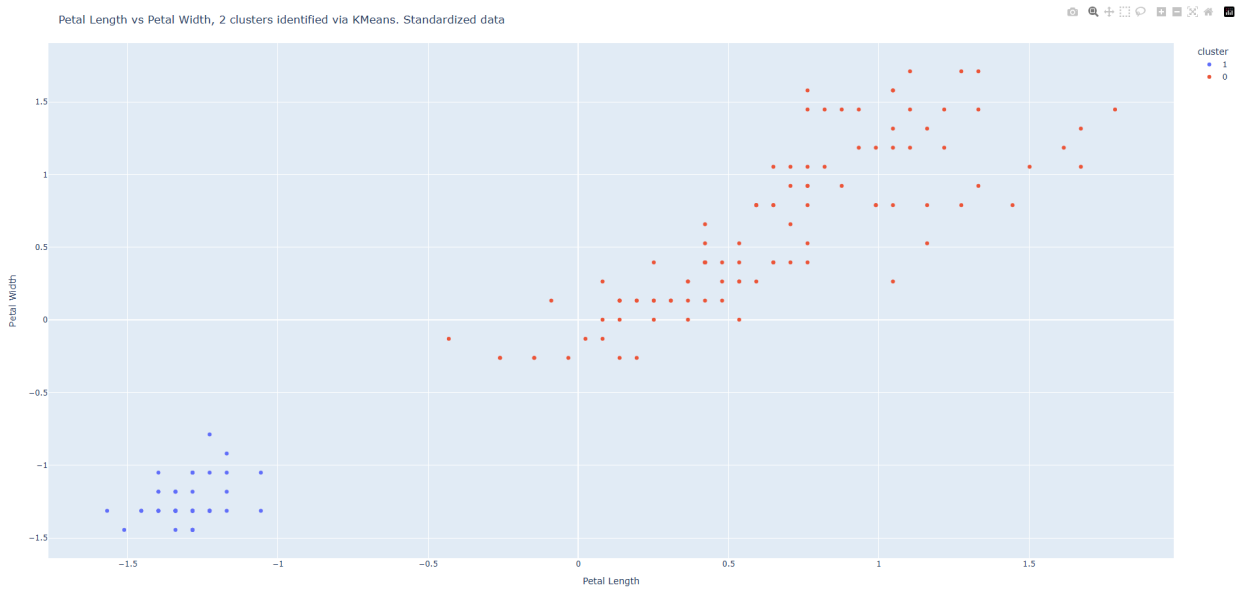


Parecen existir dos grupos claramente definidos, en la esquina inferior izquierda y superior derecha. Luego, los datos del grupo más grande podrían agruparse en clusters más pequeños posiblemente pero no se puede observar una división clara a simple vista.

2. Creen 2 “clusters” utilizando K_Means Clustering y grafiquen los resultados.

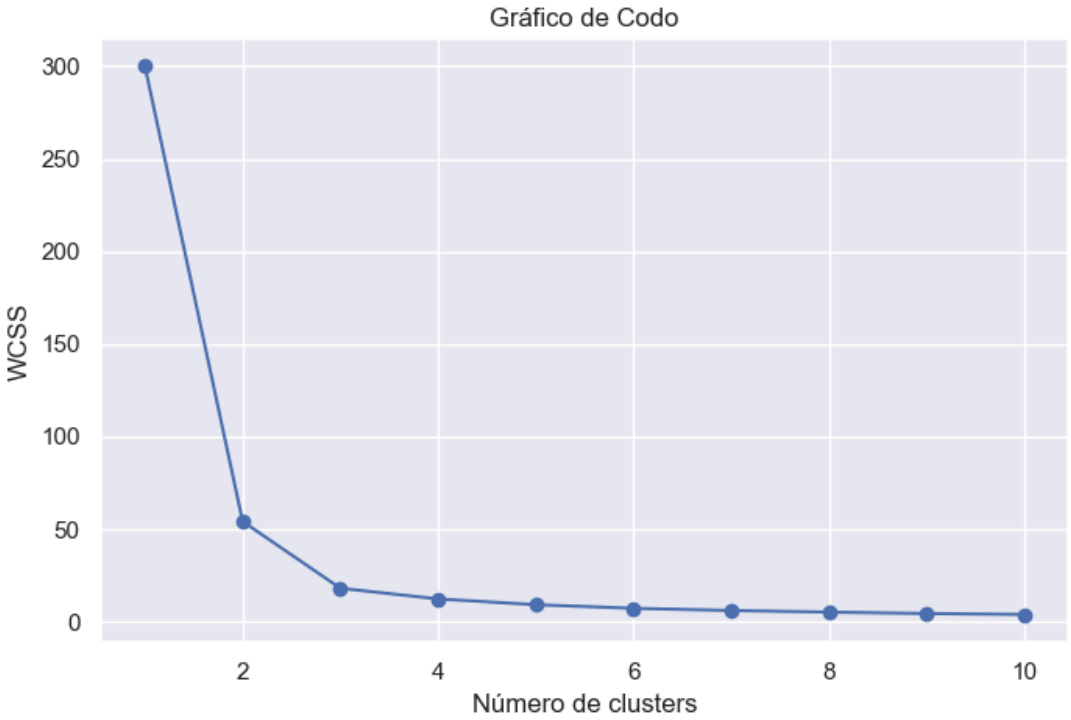


3. Estandaricen los datos e intenten el paso 2 de nuevo



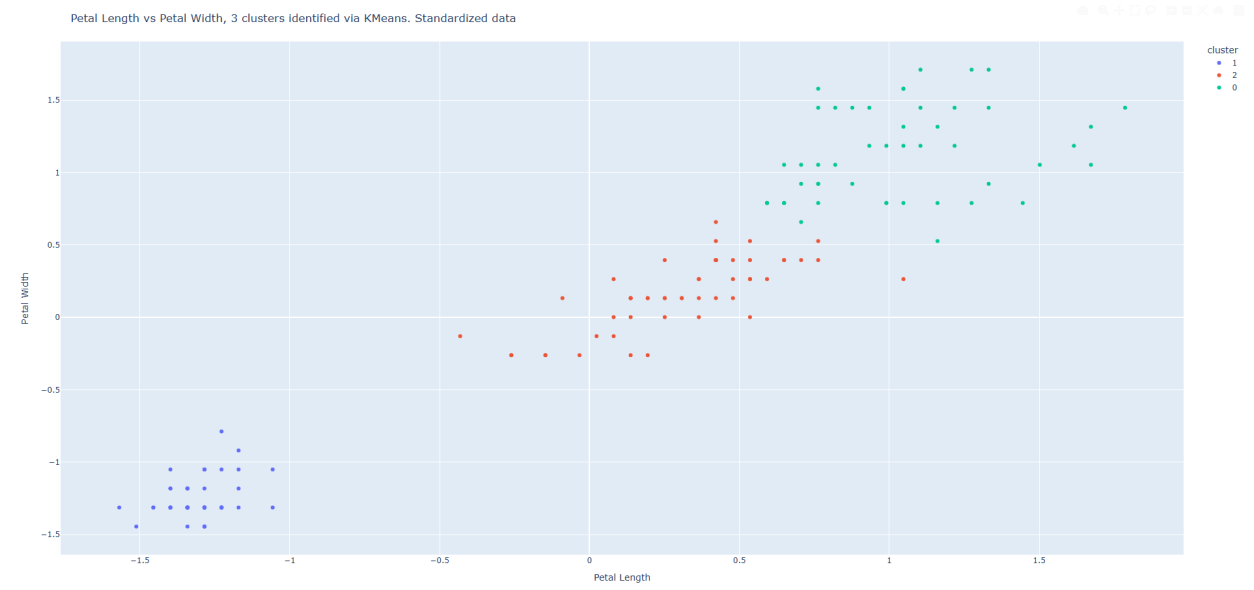
Los resultados son bastante parecidos, únicamente un punto se ve afectado por la estandarización. Esto puede deberse a que muy claramente existe una división entre ambos grupos. La pertenencia del único punto a clusters diferentes puede explicarse por el "peso" que se le da a cada variable, en este caso al estandarizar los datos se "toma más en cuenta" el ancho del pétalo y la diferencia en el eje Y de este punto respecto al cluster azul se refleja en el agrupamiento.

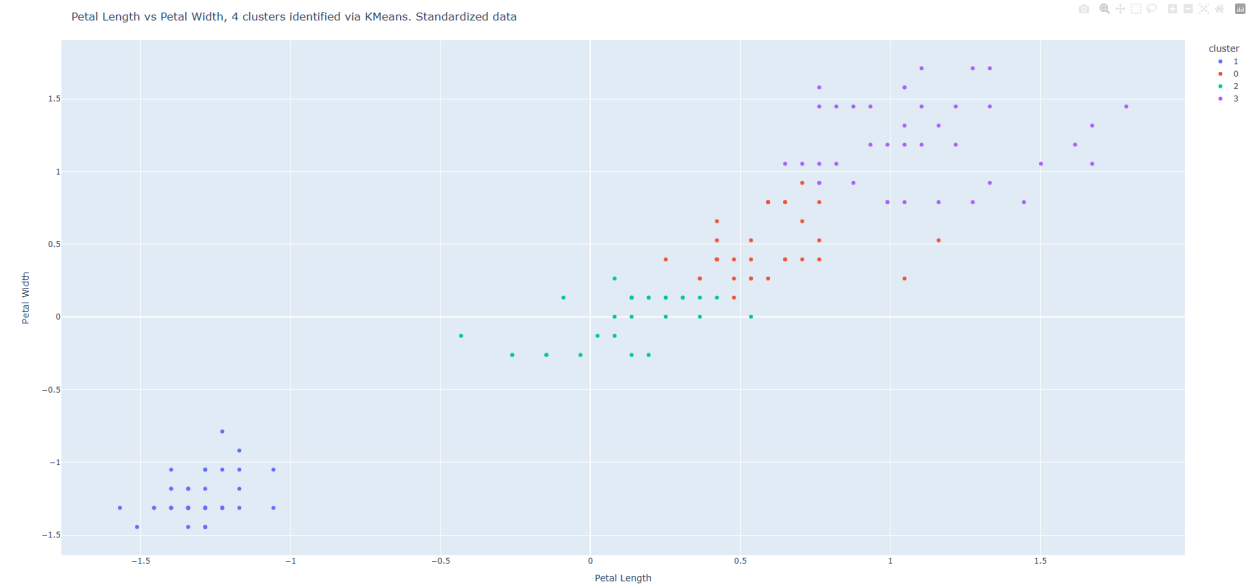
4. Utilicen el método del "codo" para determinar cuantos "clusters" es el ideal (Rango de 1 a 10)

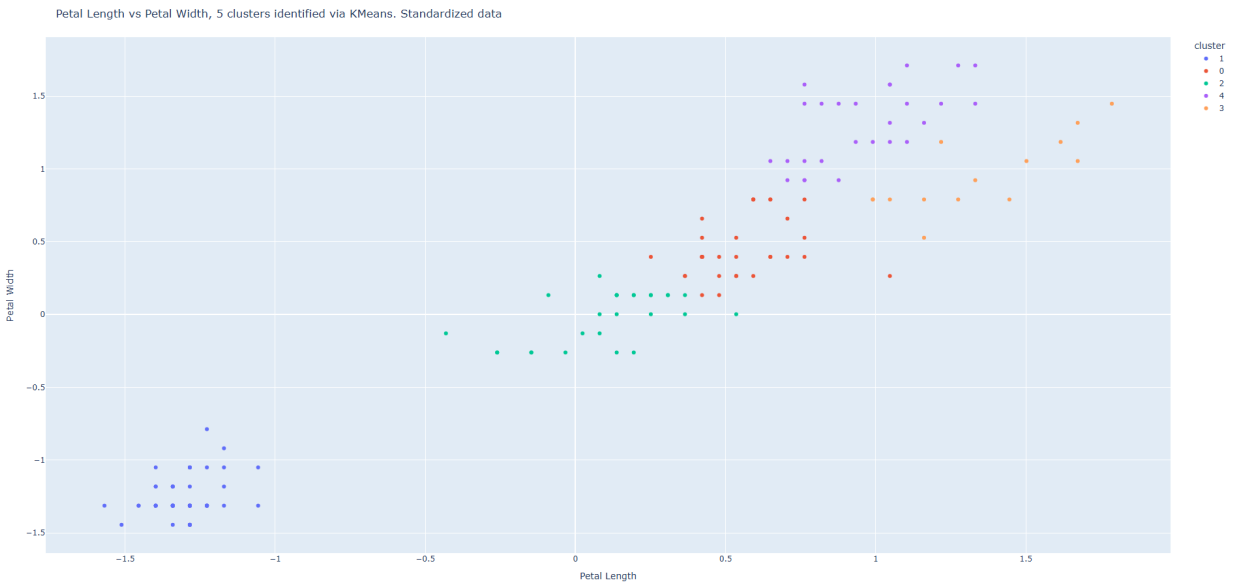


El número de clusters ideal es 3, es dónde más claramente se presenta un punto de inflexión en la gráfica.

5. Basado en la gráfica del "codo" realicen varias gráficas con el número de clusters (unos 3 o 4 diferentes) que creen mejor se ajusten a los datos

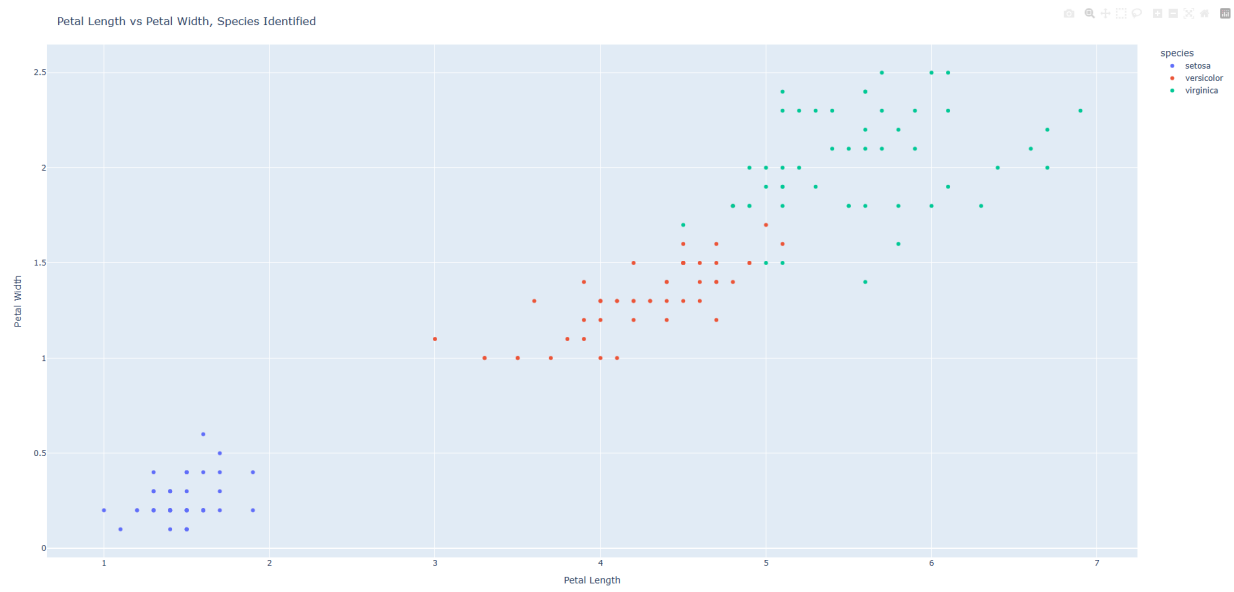

Petal Length vs Petal Width, 3 clusters identified via KMeans. Standardized data
cluster
1
2
0
Petal Width
Petal Length
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5


Petal Length vs Petal Width, 4 clusters identified via KMeans. Standardized data
cluster
1
0
2
3
Petal Width
Petal Length
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5



Elegimos los puntos 3, 4 y 5 ya que es dónde más claramente se presentan puntos de inflexión. Luego de esos puntos, la gráfica empieza a descender muy poco y no notamos una diferencia clara.

6. Comparen sus soluciones con los datos reales (Respondido en el siguiente inciso)



7. ¿Funcionó el clustering con la forma del pétalo?

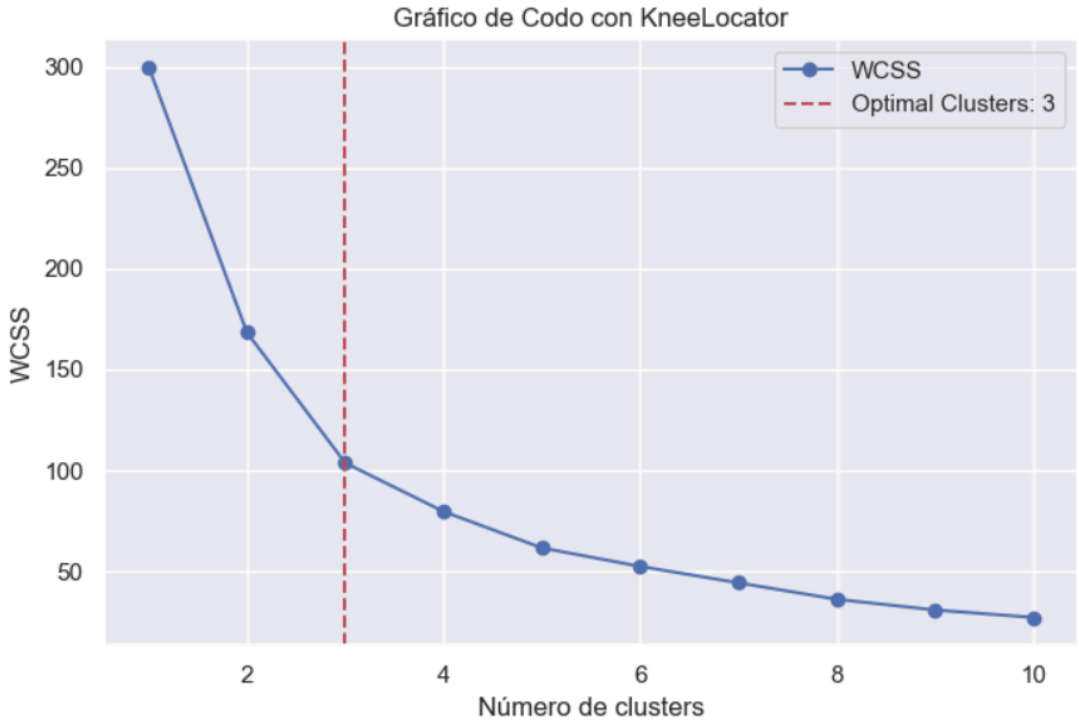
Observando las gráficas, el clustering utilizando el pétalo fue bastante acertado. En este caso, parece ser que las especies tienen una distinción más marcada en cuanto a las dimensiones / forma del pétalo que respecto al sépalo. Los principales grupos fueron identificados

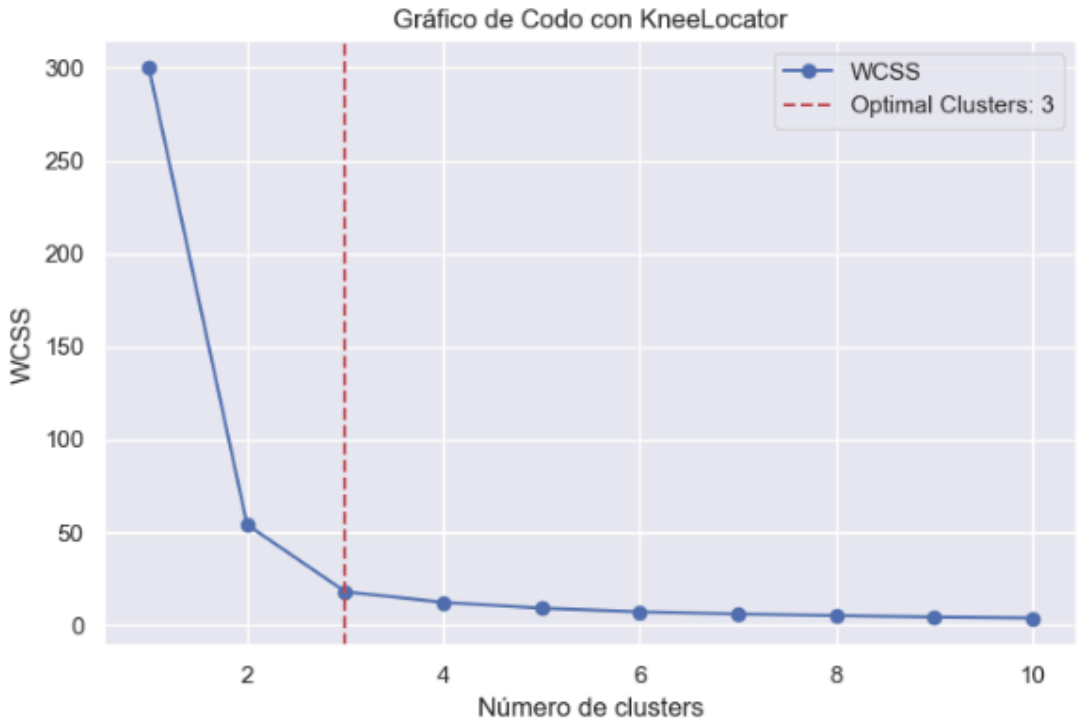
correctamente, únicamente hubo un ligero cruce entre las especies versicolor y virginica. Esto se respalda con ARI de 0.88, esto quiere indicar que los datos se encuentran muy cercanos a la realidad. En cuánto a las gráficas con diferentes números de clusters, en ninguna de ellas se logra encontrar alguna separación del grupo correspondiente a la especie setosa. Esto indica que los individuos presentes dentro de esta población tienen rasgos sumamente similares y únicos en comparación a los demás. Mientras más clusters se utilizaron, la tendencia era crear uno entre versicolor y virginica.

En conclusión, con el número adecuado de clusters el agrupamiento por forma de pétalo funcionó con la precisión esperada. Es significativamente más preciso utilizar la forma del pétalo para identificar las diferentes agrupaciones de especies dentro de este dataset.

Sección 3

Utilicen la librería "kneed" y vean si el resultado coincide con el método del "codo" que hicieron manualmente





1. ¿A que podría deberse la diferencia, si la hay?

No encontramos diferencia en este caso específico, sin embargo consideramos que es de mucha utilidad tener una herramienta que se base en un proceso matemático. En situaciones dónde no se presenten puntos de inflexión tan claros como estas gráficas, si puede existir un debate sobre el número de clusters a utilizar en el análisis.

2. ¿Les dió el número correcto de clusters, comparado a los datos reales?

Si, nos dio el número de clusters correctos.

3. Basado en los resultado que tuvieron, ¿A qué conclusiones llegaron?

- Las especies presentan una mayor distinción entre una y otra en cuánto a dimensiones del pétalo, haciendo este atributo más apropiado para la técnica de clustering.

- La estandarización de los datos es sumamente importante, en especial en atributos tales como la forma del sépalo en este caso. Los agrupamientos pueden tomar rutas completamente diferentes al enfrentarse con variables numéricamente mayores a las demás.

- Este algoritmo presenta limitaciones en cuanto a la correcta segmentación cuando entre agrupaciones se tienen atributos similares, o “overlapping”. Si un mismo individuo de una especie se puede encontrar ya sea en el límite superior, o el inferior dentro de una muestra resulta poco probable que sea agrupado utilizando K-Means.

- Es importante encontrar el número correcto de clusters para un dataset, ya que de lo contrario las agrupaciones pueden dividirse en dos o combinar datos de diferentes categorías.