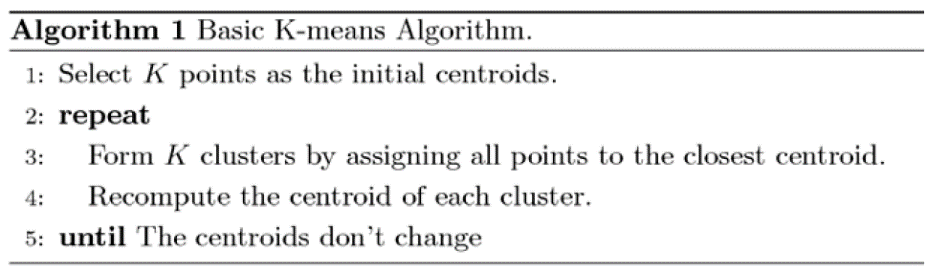
Apellido y nombre: Acero Mamani Washington

Codigo:141850

**Algoritmo K-means**

Después de probar el algoritmo k-means en una muestra de datos pequeña como 15 datos, el centroide se encuentra en 2 iteraciones, y el centroide no se mueve más, pareciera que el algoritmo no funciona(esto al programarlo, parece como si el algoritmo no funcionase o que está mal programado, a diferencia de lo que sucede con las librerías que implementan el algoritmo), pero al probar con una cantidad de datos razonable como 3000 se nota claramente que el centroide se mueve, en las primeras iteraciones varía mucho y poco a poco ya no tanto y se obtiene los centroides deseados.

**Algoritmo K-means**



Implementacion del algoritmo:

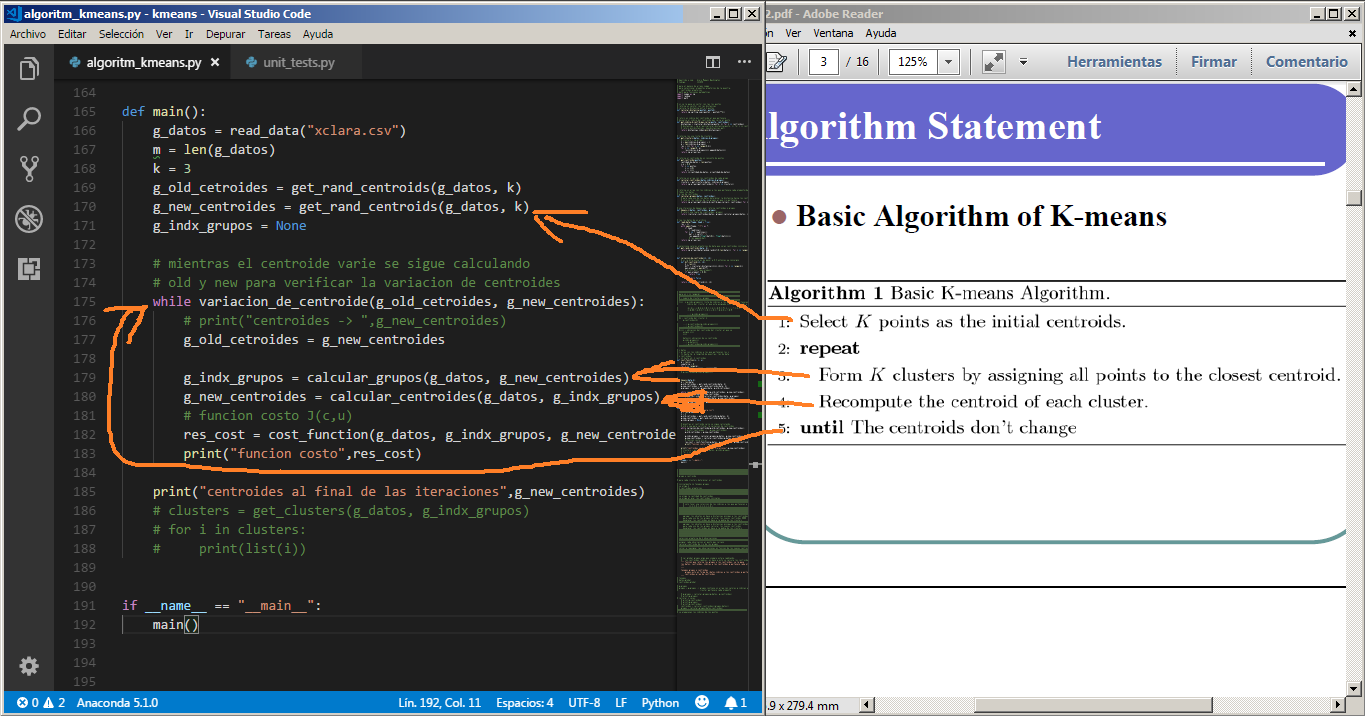


Imagen de el algoritmo en pseudocidgo y la versión implementada en Python.

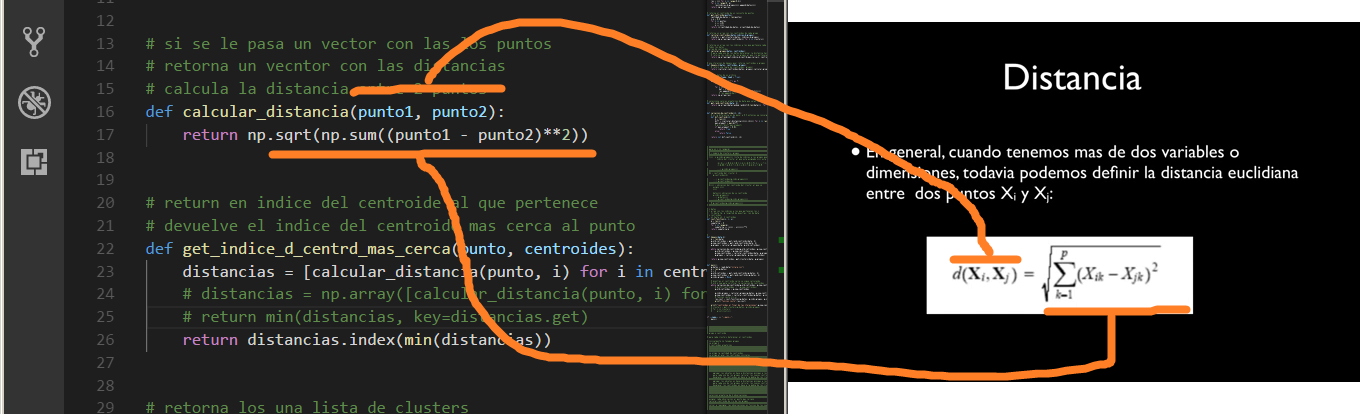
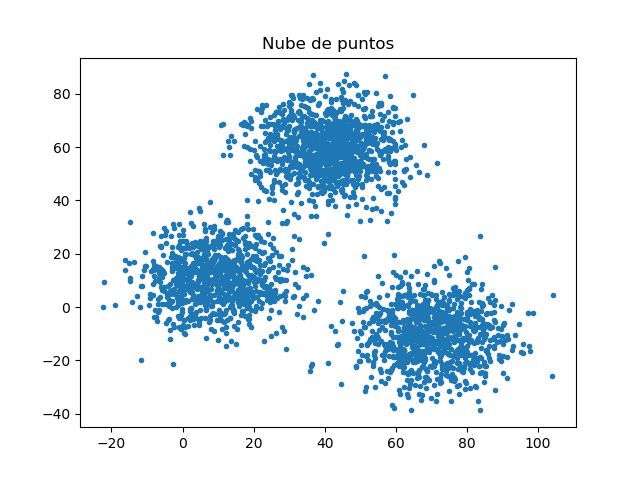


Imagen de la implementación de la función distancia.

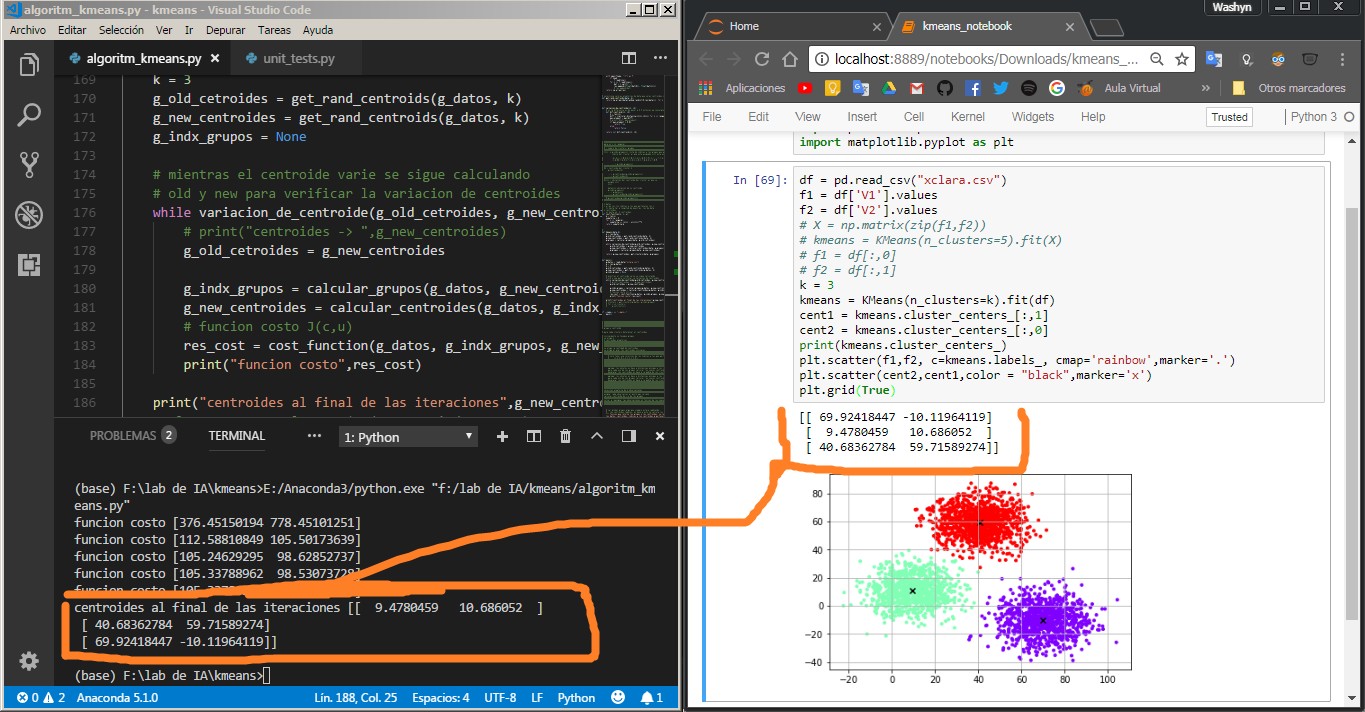
**Pruebas:**

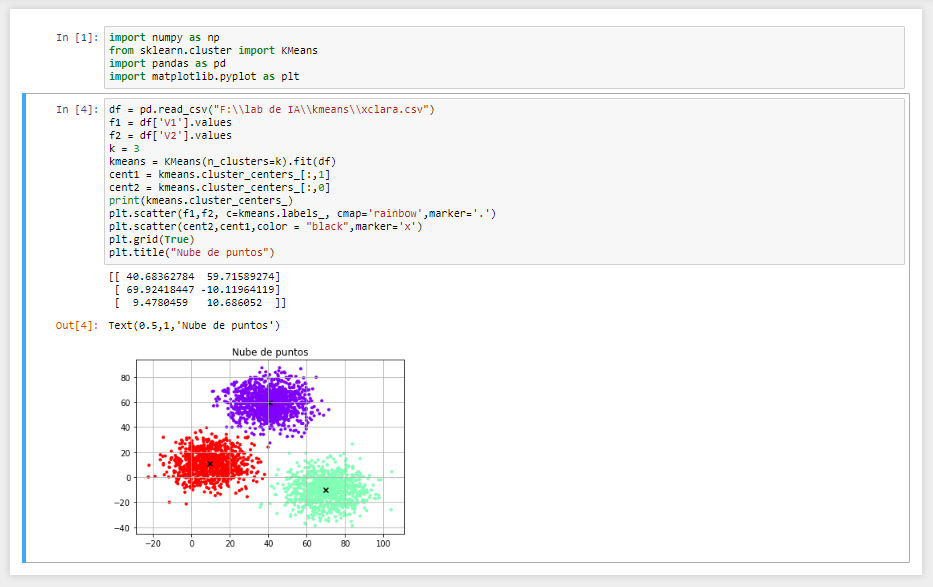
Para realizara las pruebas se usó los datos xclara.csv, se puede observar que los centroides al final son los mismos al usar la librería y la versión implementada, pero esto no sucede cuando se tiene una muestra de datos m muy pequeña o cuando los clusters no se pueden notar a simple vista en esos casos se ve una diferencia clara.

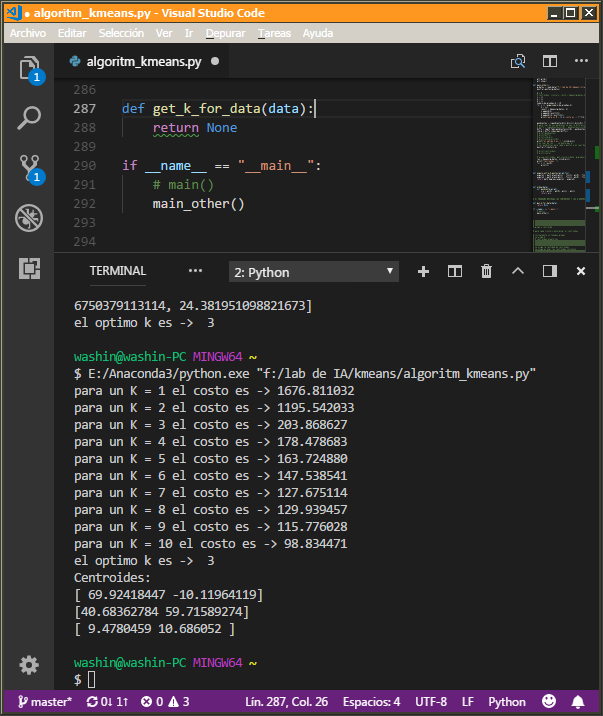
Para realizar las pruebas se uso un set de datos xclara.csv que se puede encontrar en internet para este tipo de pruebas, contiene 2 columnas, y 3000 muestras.



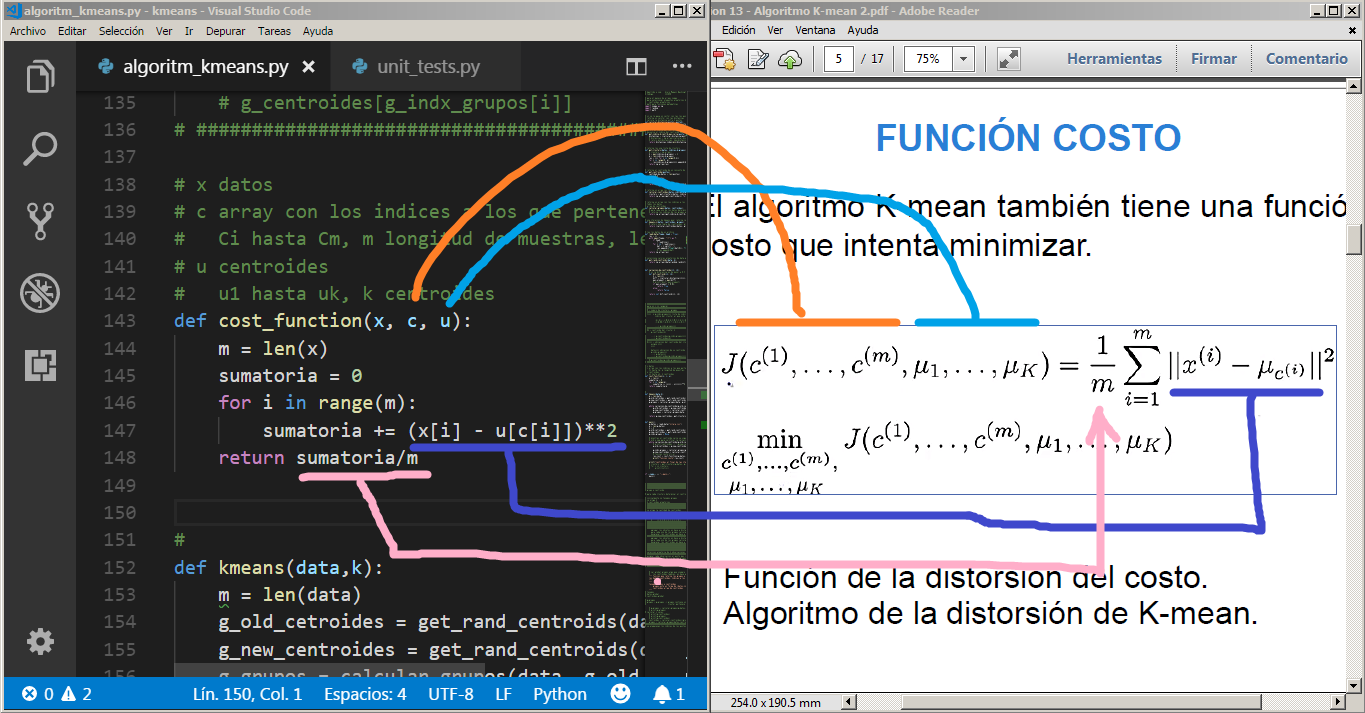
Comparación de los valores de centroides obtenidos usando la librería vs el algoritmo implementado.



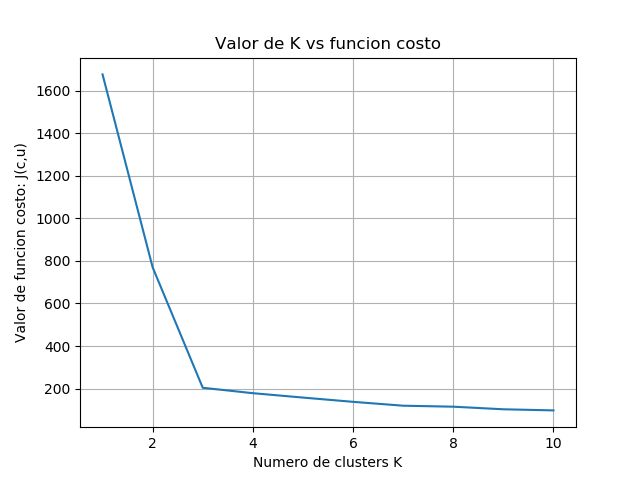




**Implementación de la función costo:**



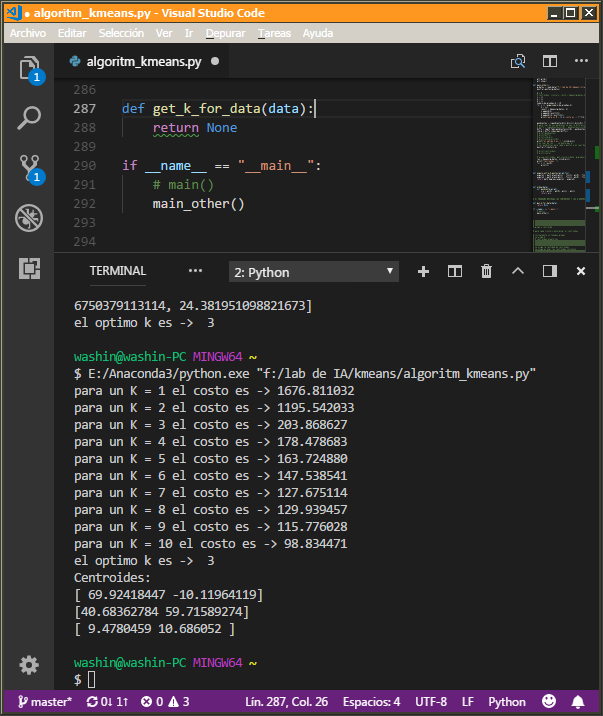
**Resultados de la función costo:**



Grafica de la funcion costo:

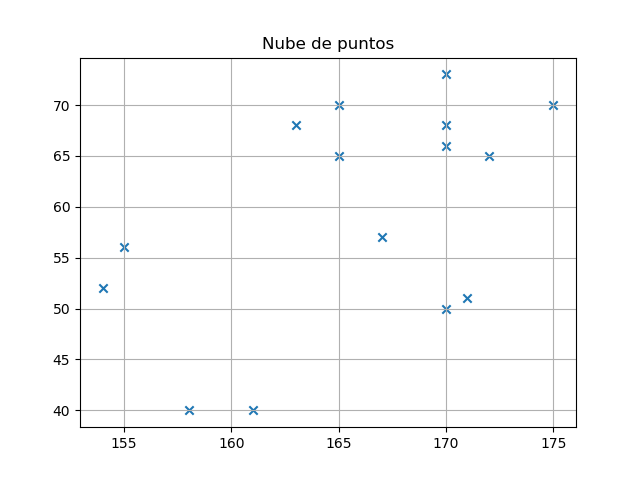
Para implementar el Metodo de elbow se tomo en cuenta la pendiente de entre 2 puntos, de esto se puede ver en cual de los 2 puntos la pendiente tiene una mayor variacion, para este caso se observa que hay una diferencia entre k[2], k[3] y k[3], k[4] entonces de pude decir que el k adecuado esta ahí.

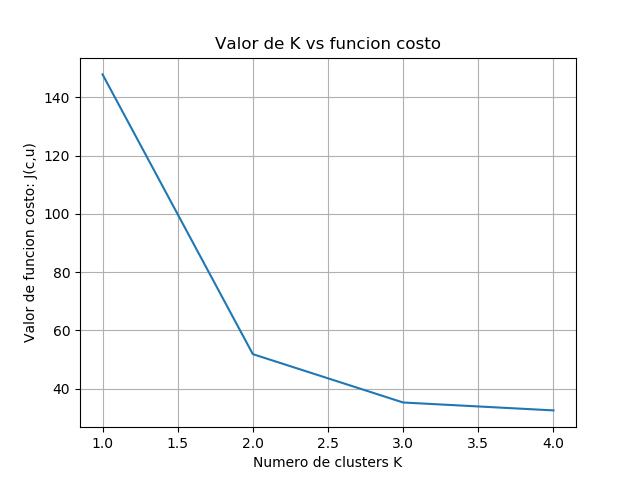
Entonces el algoritmo nos da el valor de k(cantidad de clusters que debemos consiferar).

****

Observaciones: La ejecución de la función costo aumenta a medida que se incrementa el valor de K cuando hay bastantes datos.

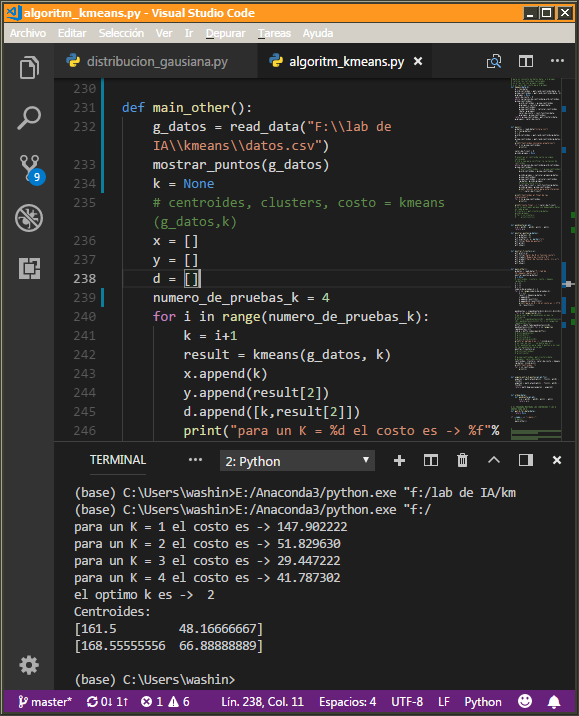
**Prueba con los datos obtenidos en clase Talla vs Peso:**





Este grafico nos muestra que el valor adecuado para k podría ser 2.

Resultados obtenidos al ejecutar el algoritmo sobre los datos talla vs peso.



Para el valor k = 2 los centroides obtenidos se muestran en el siguiente grafico en base a la ejecución del algoritmo.

