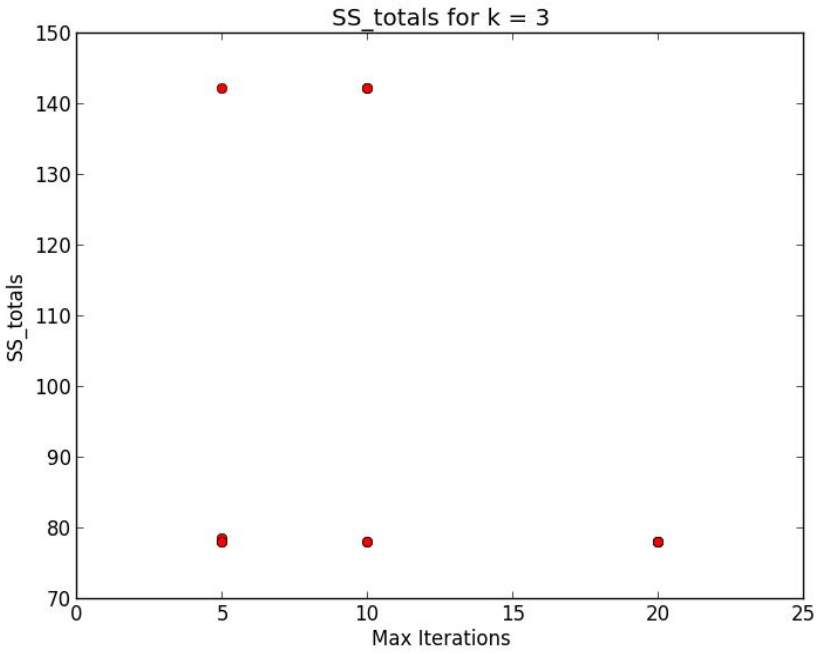
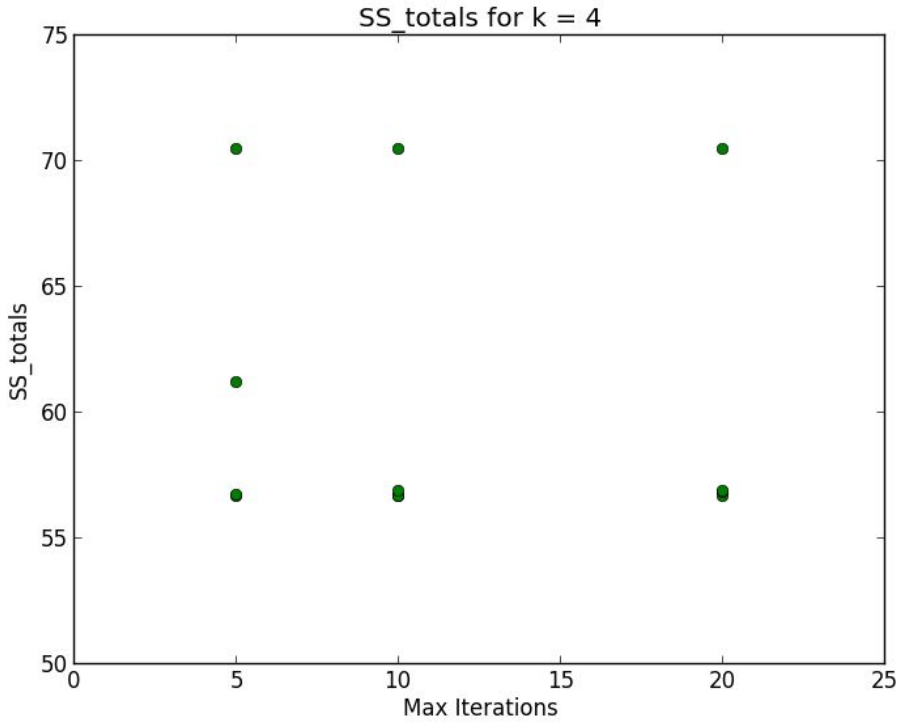
Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

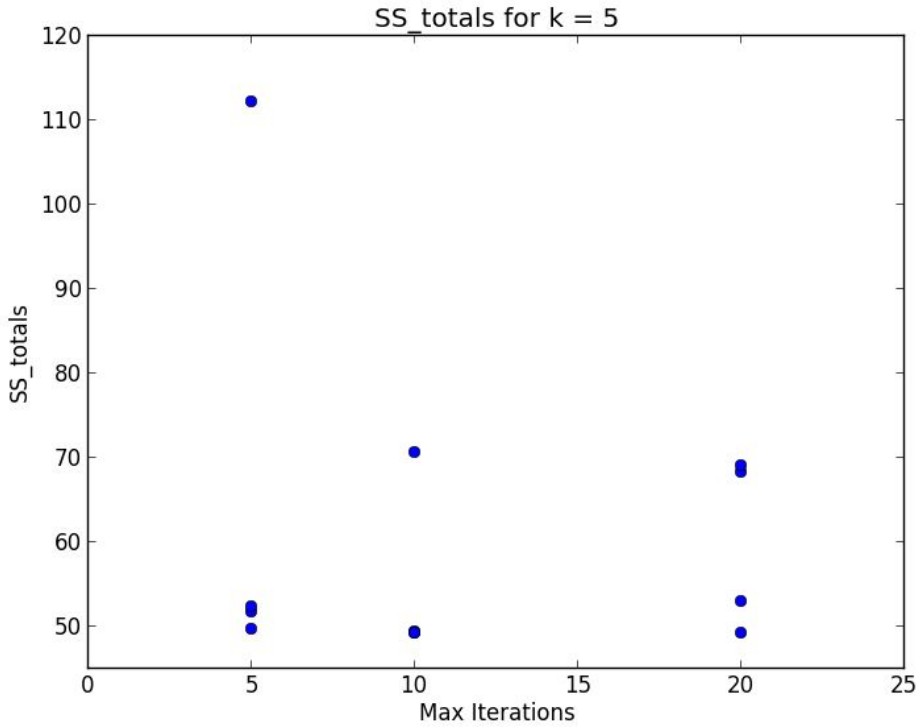
Part A

The next 3 graphs show the 4 SS_totals computed per each choice of max iterations and k. In many instances, there is one outlier with a very large value which makes the other more realistic data points overlap each other and appear as a single point. But know that for every max iter, there are 4 data points. My raw data is also available in data/out1.txt

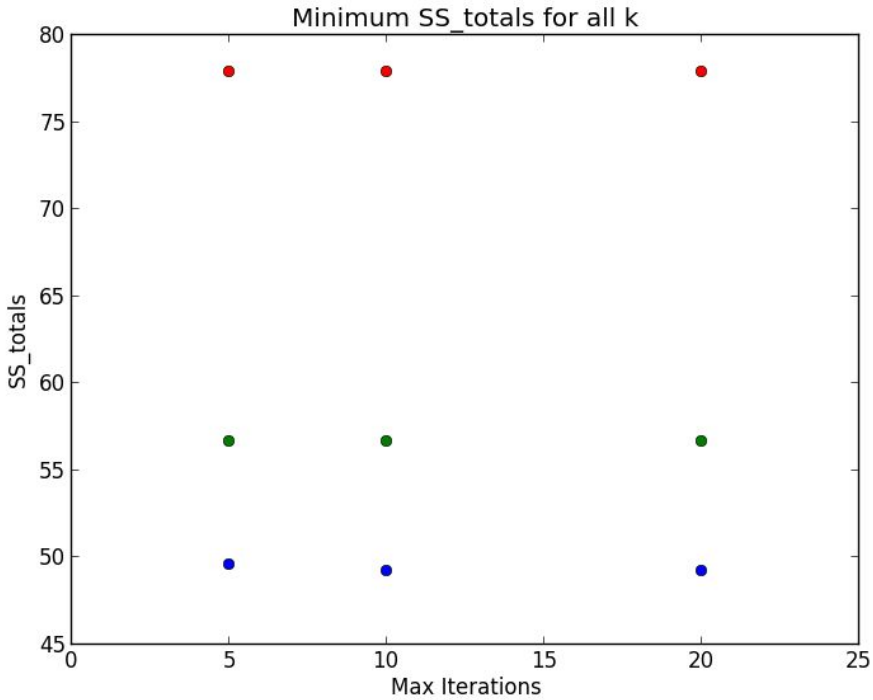




Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

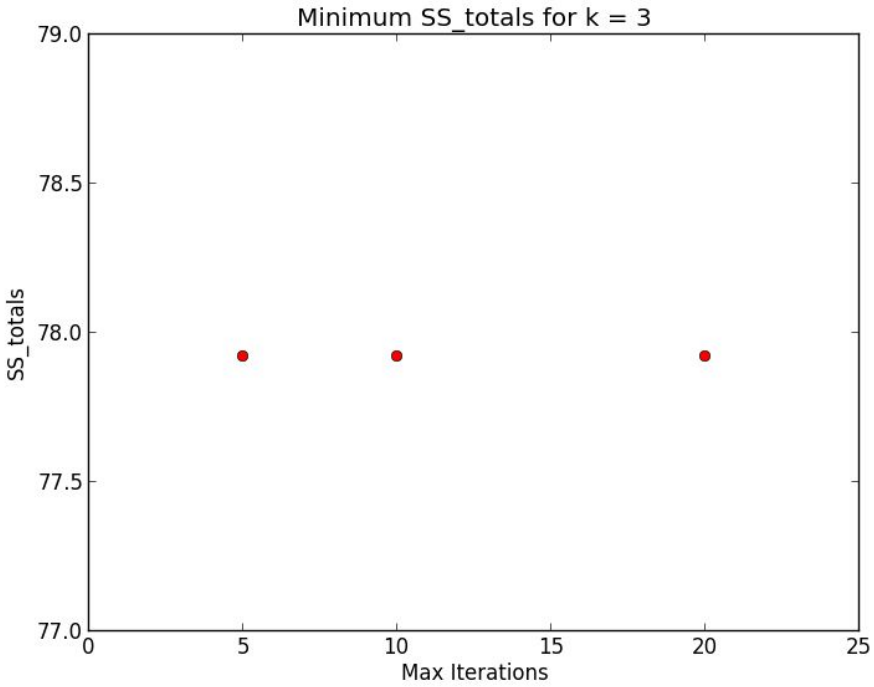


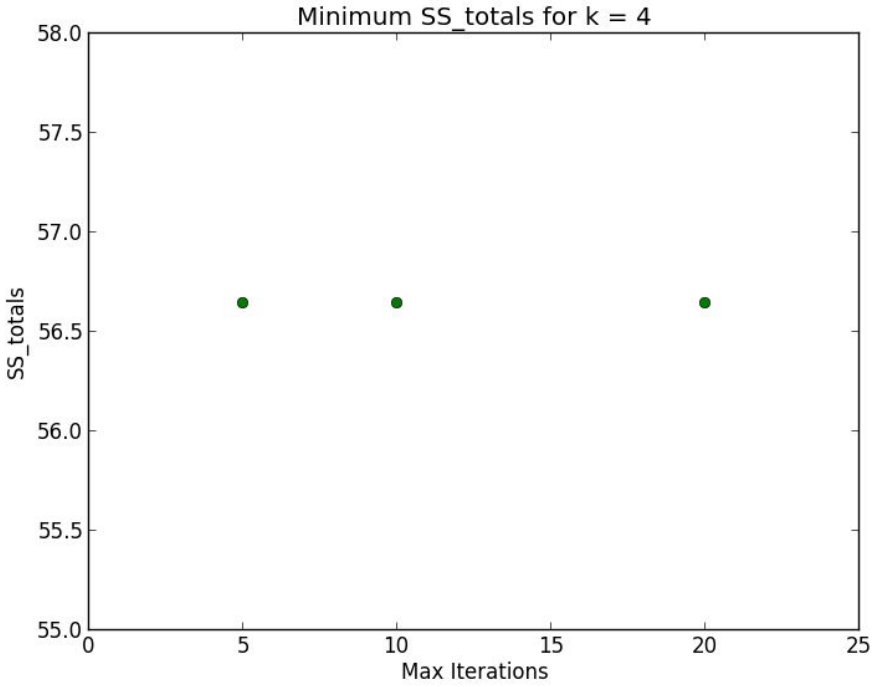
The next graph plots the minimum values from every run.



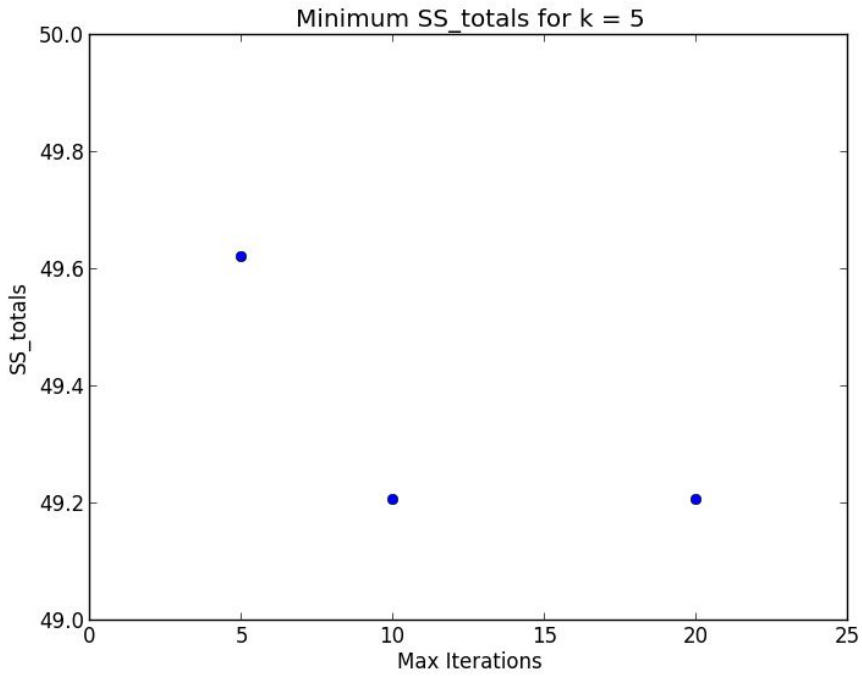
Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

The next three graphs are the minimum totals split by k to better represent the variation in the plots with change in max iter. But this variation is only evident in the k=5 graph.





Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8



My first takeaway is that randomization and rerunning for the same hyperparameters is essential. There was such a massive variability in the data. For example, I had the following data. 2 out of 4 times I got a bad value like 142.

1. k: 3, max iter: 10, SS_total: 77.9198903509
2. k: 3, max iter: 10, SS_total: 142.103741289
3. k: 3, max iter: 10, SS_total: 77.9198903509
4. k: 3, max iter: 10, SS_total: 142.187677109

A larger number of max iterations leads to lower SS_totals, although this is a marginal difference and only noticeable in my raw data, not in my graphs.
Our choice of k also seriously influences SS_totals. More clusters would mean data points are more likely to be closer to their assigned cluster center, so this will reduce the SS_total, as the data shows.

I don't think we should let a preference for low SS_totals influence choice in k. As stated earlier, a greater k will lead to smaller SS_total. We can think of N clusters for N examples, where each example gets its own cluster. This effectively makes the SS_total 0. There are other metrics as we will see in part B which can help us choose a better k.

Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

Part B

Just by eyeing my data, presented in *data/out2.txt*, I decided that choosing the optimal cluster by the maximal number of the iris type in a cluster will always work correctly.
Then I computed my F1 scores, shown below. They are also presented in *data/total_f1.txt*. I inserted my averaged F1 scores as a textbox on the side.

k	*maxiter*	*Iris-setosa*	*Iris-virginica*	*Iris-versicolor*	Averaged
3	*5*	*1.0*	*0.827586206897*	*0.864864864865*	0.896564403954
3	*10*	*1.0*	*0.827586206897*	*0.864864864865*	0.896564403954
3	*20*	*1.0*	*0.827586206897*	*0.864864864865*	0.896564403954
k	*maxiter*	*Iris-setosa*	*Iris-virginica*	*Iris-versicolor*	
4	*5*	*1.0*	*0.79012345679*	*0.692307692308*	0.825383904749
4	*10*	*1.0*	*0.79012345679*	*0.692307692308*	0.825383904749
4	*20*	*1.0*	*0.79012345679*	*0.692307692308*	0.825383904749
k	*maxiter*	*Iris-setosa*	*Iris-virginica*	*Iris-versicolor*	0.678486372624
5	*5*	*0.6087*	*0.80487804878*	*0.621621621622*	0.73395533332
5	*10*	*0.72*	*0.79012345679*	*0.692307692308*	0.73395533332
5	*20*	*0.72*	*0.79012345679*	*0.692307692308*	

As I mentioned in Part A, the SS_total is not a reliable metric for choosing the k value. The F1 scores tell us to choose k=3, max iter = 20. This answer makes sense when we have 3 labels in the data and more runs helps to normalize the randomized initializations in my runs.

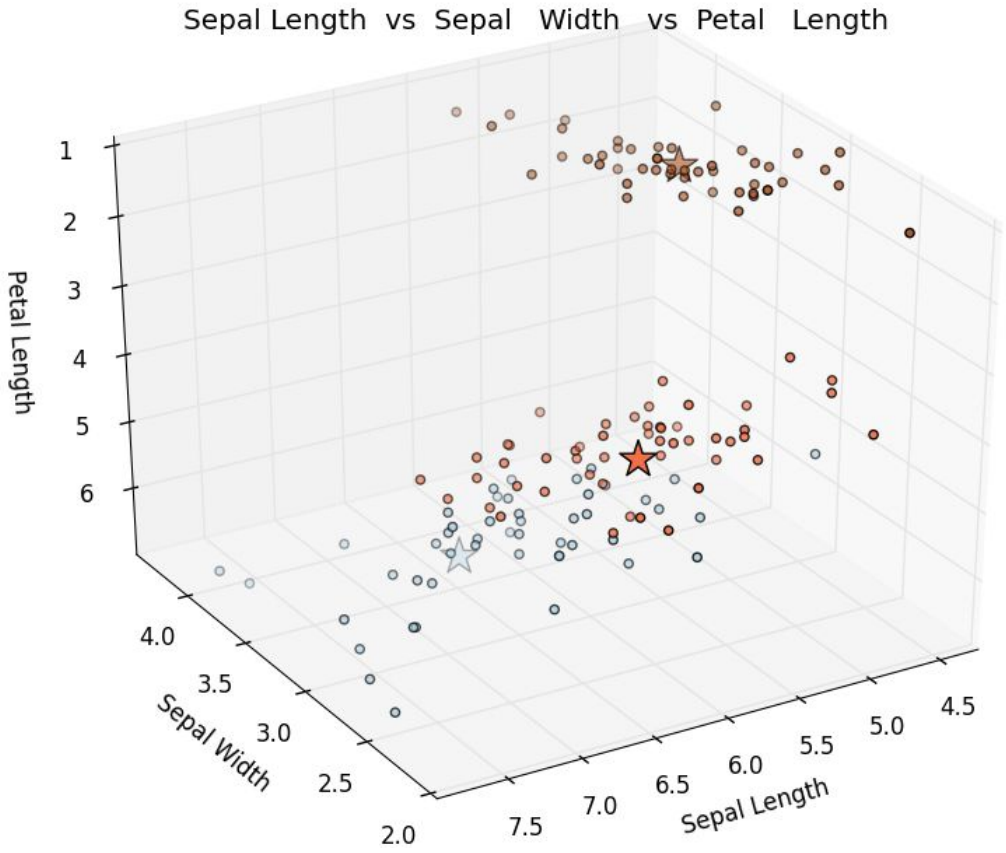
Following this reasoning, my choice for the best cluster was k=3, max iter = 20. A breakdown of the cluster composition and central coordinates is shown below.

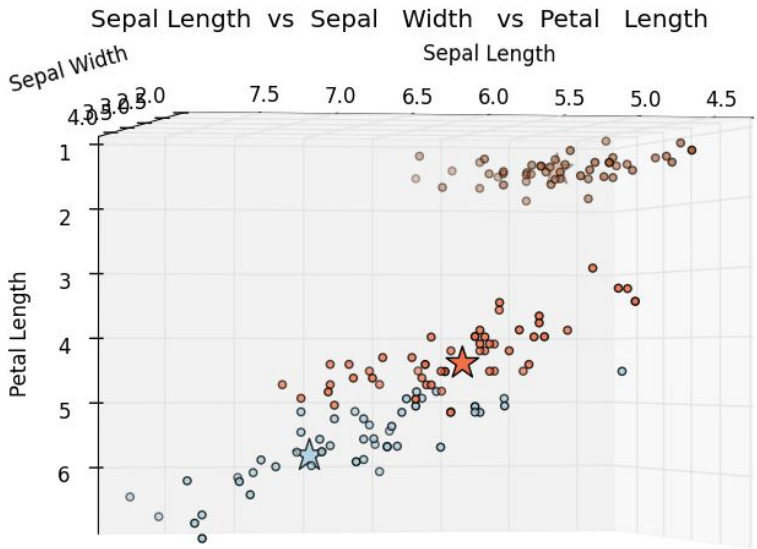
{'Iris-virginica': 36, 'Iris-versicolor': 2}	(6.84999, 3.07368, 5.74211, 2.07105)
{'Iris-setosa': 48}	(5.01042, 3.43125, 1.46250, 0.24999)
{'Iris-virginica': 13, 'Iris-versicolor': 48}	(5.90328, 2.74918, 4.38197, 1.42623)

The next section(on the next page) is my 3d graph plots. I include two views of each graph to better demonstrate the clustering.

Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

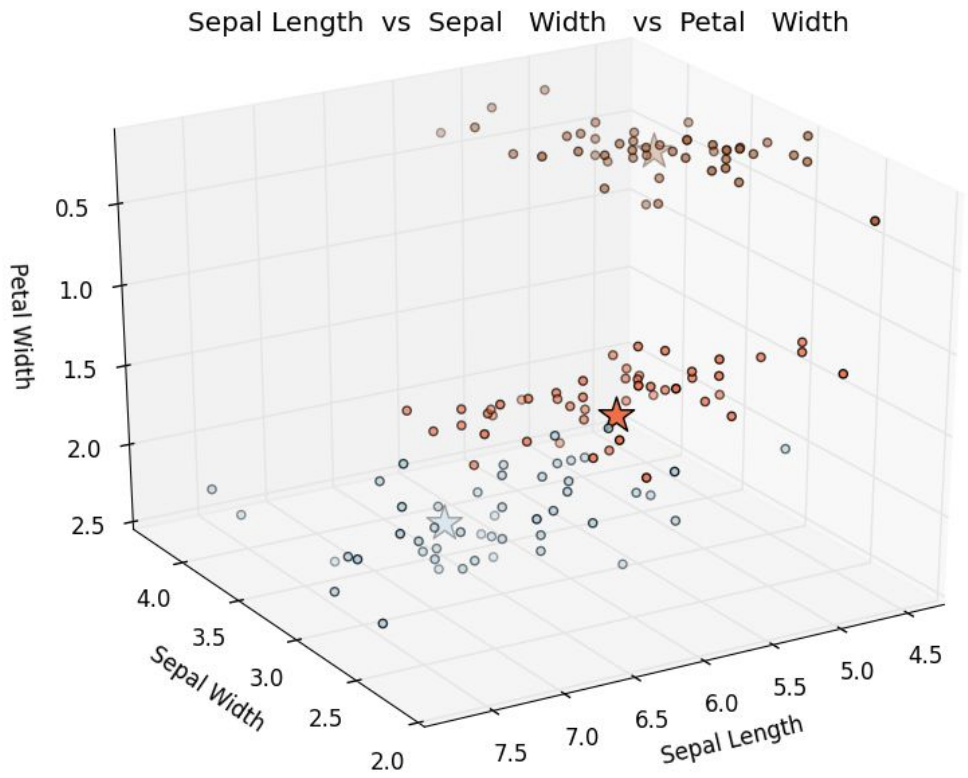
a) Sepal Length vs Sepal Width vs Petal Length

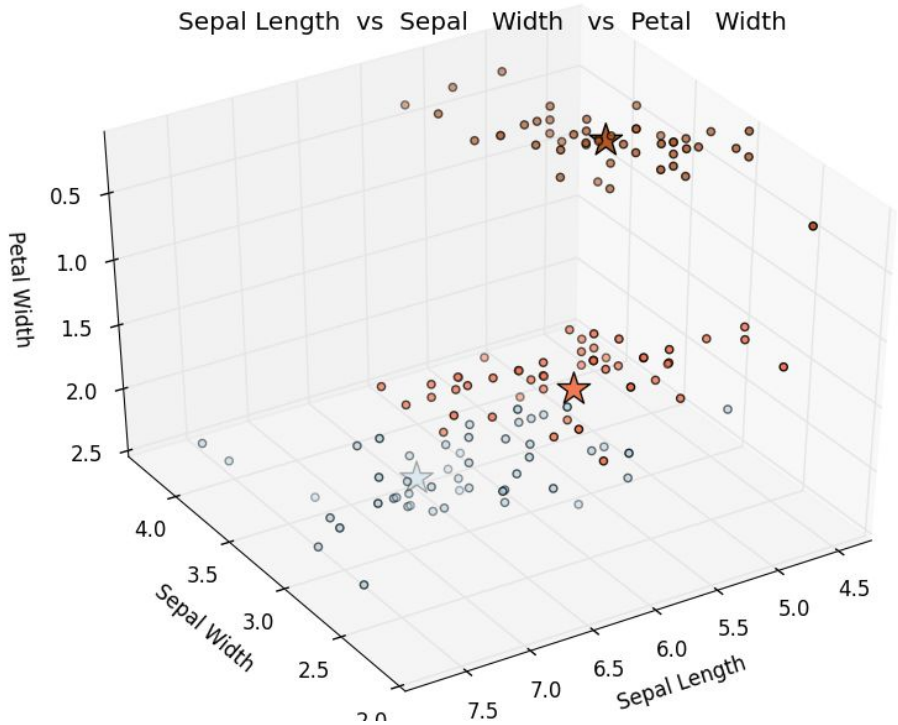




Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

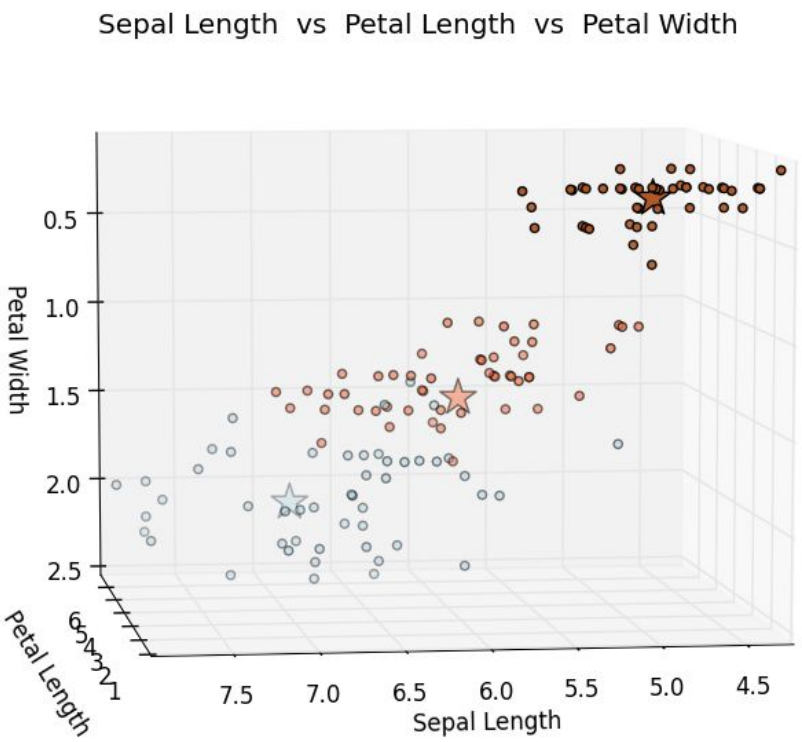
b) Sepal Length vs Sepal Width vs Petal Width

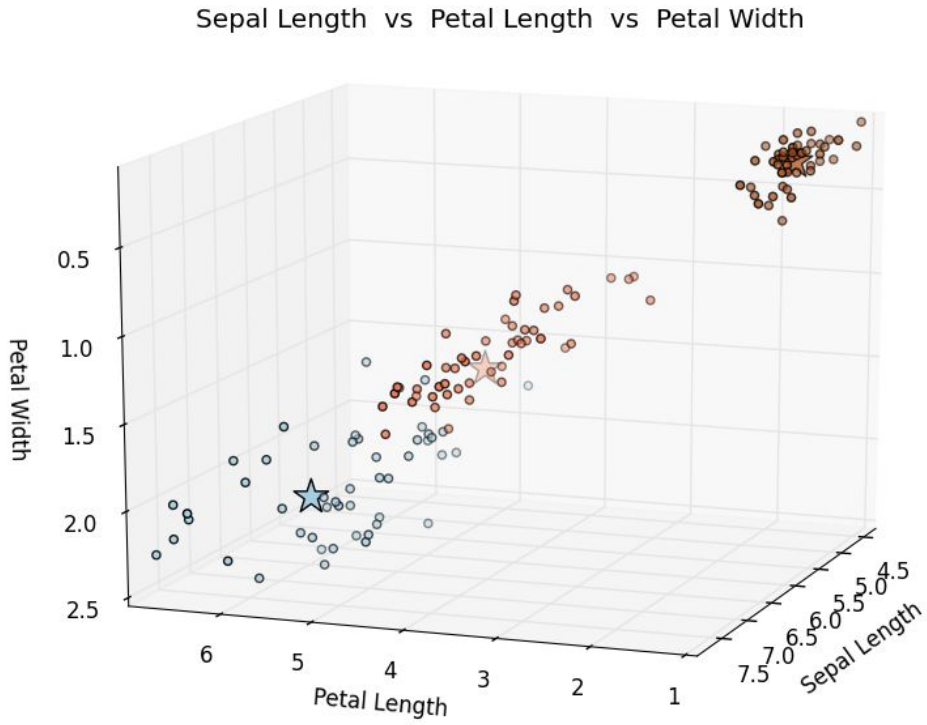




Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

c) Sepal Length vs Petal Length vs Petal Width





Ranvit Bommineni
bommine2@illinois.edu
CS 440 HW 8

d) Sepal Width vs Petal Length vs Petal Width