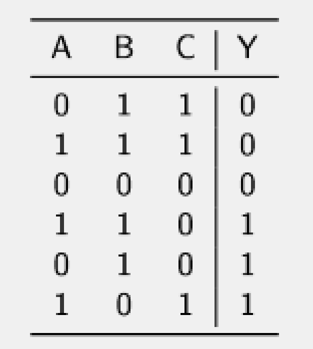
Jack Stevenson
CS 434H
29 Nov 2022

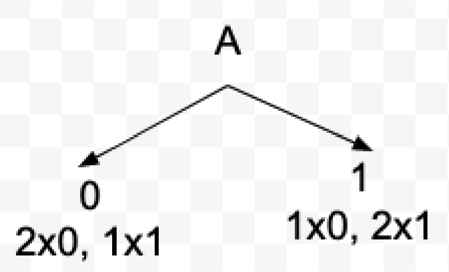
1: Manually learning a decision tree



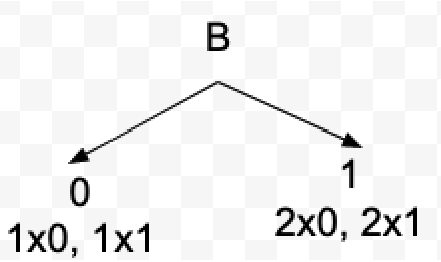
A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

$P(Y = 0) = 1/2$
$P(Y = 1) = 1/2$
$H(Y) = -1/2log_2 1/2 - 1/2log_2 1/2 = 1$

We can start by trying to split on individual variables. However, we notice that there is not a variable correlated with the label Y:

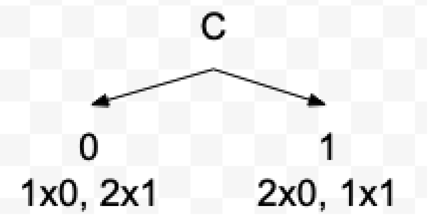


$P(Y = 0|A = 0) = 2/3$
$P(Y = 1|A = 0) = 1/3$
$P(Y = 0|A = 1) = 1/3$
$P(Y = 1|A = 1) = 2/3$
$H(Y|X) = -1/2(2/3log_2 2/3 + 1/3log_2 1/3) - 1/2(1/3log_2 1/3 + 2/3log_2 2/3) = 0.918$



$P(Y = 0|B = 0) = 1/2$
$P(Y = 1|B = 0) = 1/2$
$P(Y = 0|B = 1) = 1/2$

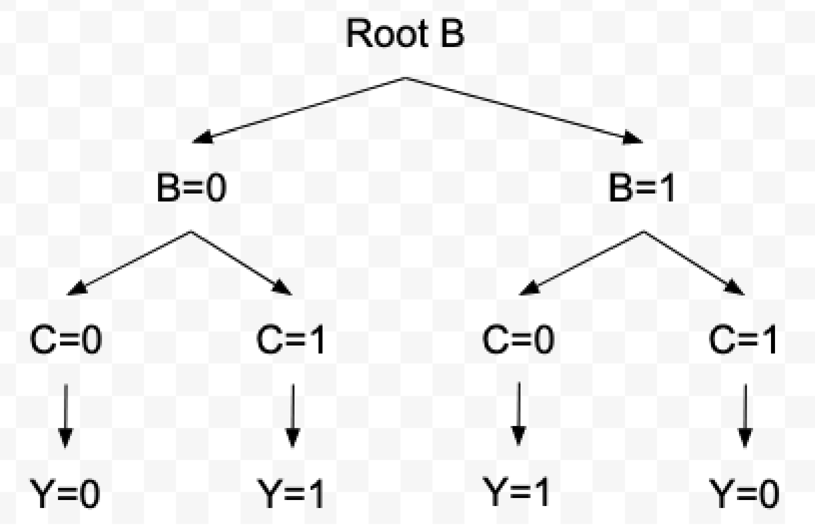
$P(Y=1|B=1)=1/2$
$H(Y|X)=-1/2(1/2log_2 1/2+1/2log_2 1/2)-1/2(1/2log_2 1/2+1/2log_2 1/2)=1$



$P(Y=0|C=0)=1/3$
$P(Y=1|C=0)=2/3$
$P(Y=0|C=1)=2/3$
$P(Y=1|C=1)=1/3$
$H(Y|X)=-1/2(1/3log_2 1/3+2/3log_2 2/3)-1/2(2/3log_2 2/3+1/3log_2 1/3)=0.918$

We reduce entropy by splitting on A or C. This results in an information gain of 1 - 0.918 or 0.082.

We can now consider multiple splits. After inspecting the data, it is apparent that Y is 0 when B and C match and is 1 when they do not match. Let us construct a decision tree that follows this observation:



$P(Y=0|B=0,C=0)=1$
$P(Y=1|B=0,C=1)=1$
$P(Y=1|B=1,C=0)=1$
$P(Y=0|B=1,C=1)=1$
$H(Y|X)=-1/2(0+0)-1/2(0+0)=0$

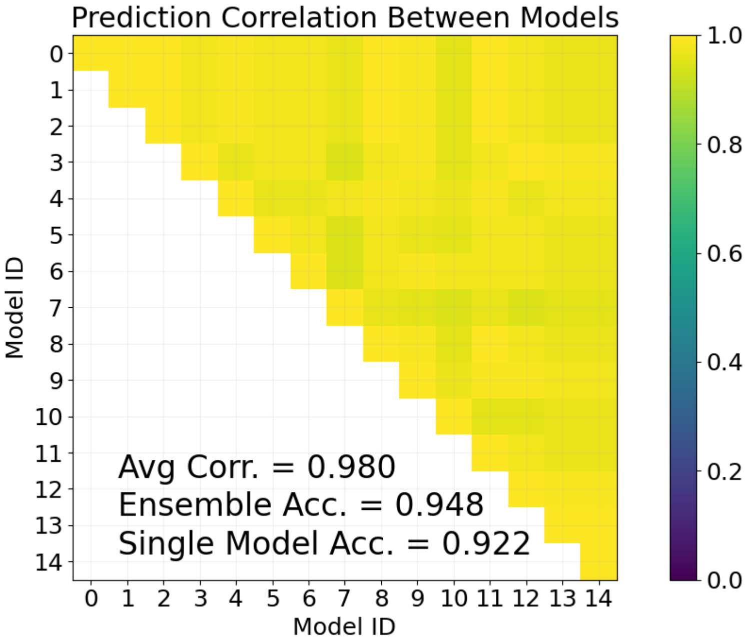
$IG(B,C)=H(Y)-H(Y|B,C)=1-0=1$

We can confirm that this tree correctly maps the input to the output:
$B=1,C=1\rightarrow Y=0$
$B=1,C=1\rightarrow Y=0$

$B = 0, C = 0 \to Y = 0$
$B = 1, C = 0 \to Y = 1$
$B = 1, C = 0 \to Y = 1$
$B = 0, C = 1 \to Y = 1$

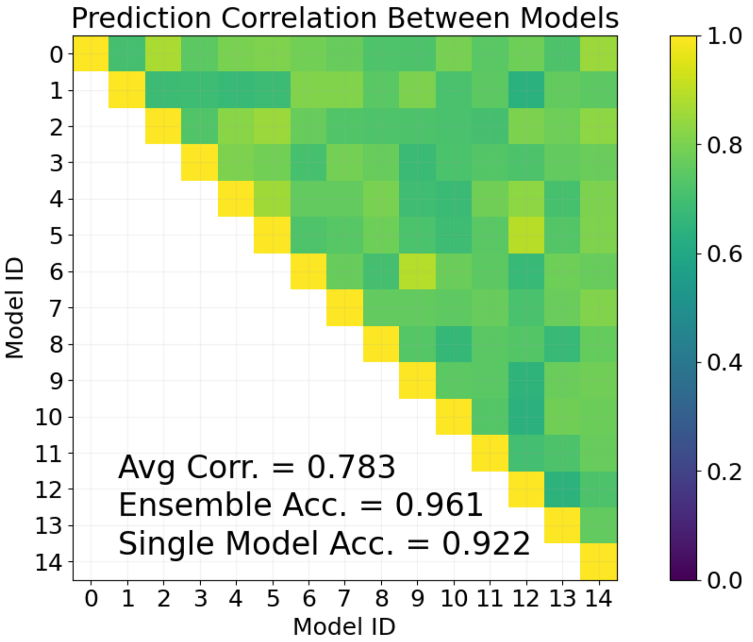
2: Measuring correlation in random forests

After applying bagging to the sampling of training datapoints, I noticed that ensemble accuracy increased by around 2.5%.



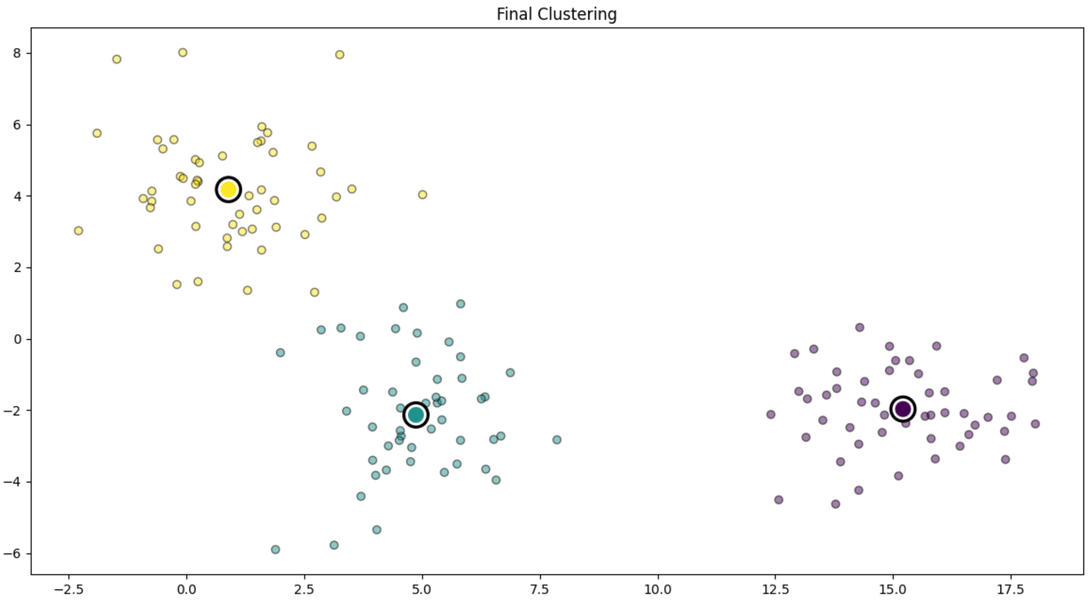
Correlation plot with bagging implemented

I set the maximum number of features to 10. This drastically reduced the average correctness of a single classifier, but it led to a significant increase in ensemble accuracy of almost 4%.



Correlation plot with a max_features value of 10

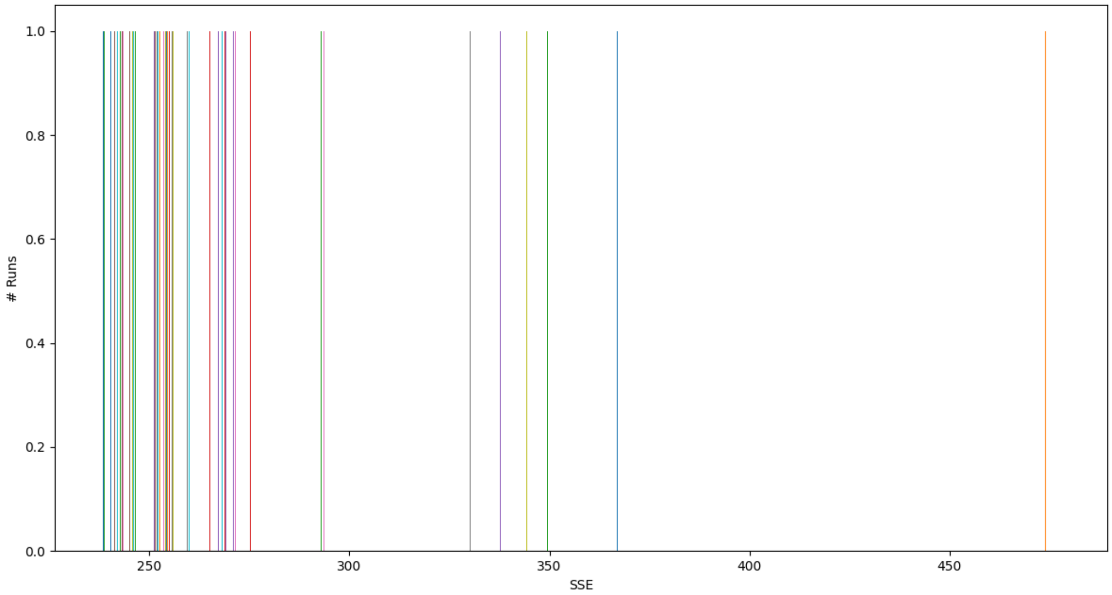
3: Implement k-means



Visualization plot of clusters

4: Randomness in clustering

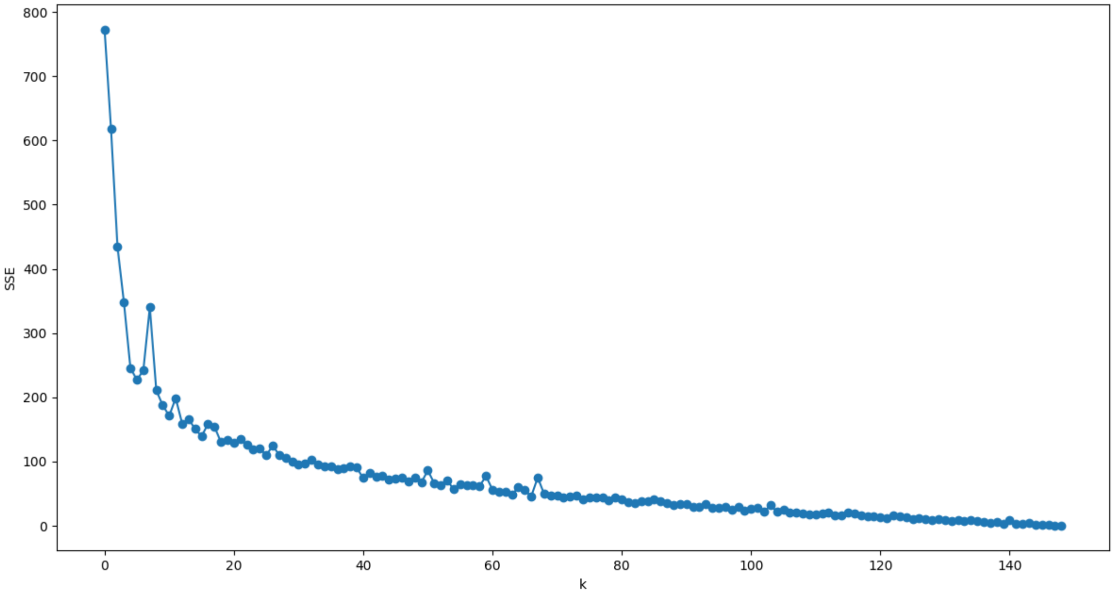
The SSE can vary greatly between runs, and the largest SSE is more than double the lowest. When applying k-means to a real dataset, it would be valuable to run multiple trials. Then, with the number of centroids constant, you could find the clustering pattern with the lowest SSE and have some confidence that it is a good model.



SSE plot for 50 runs at k = 5

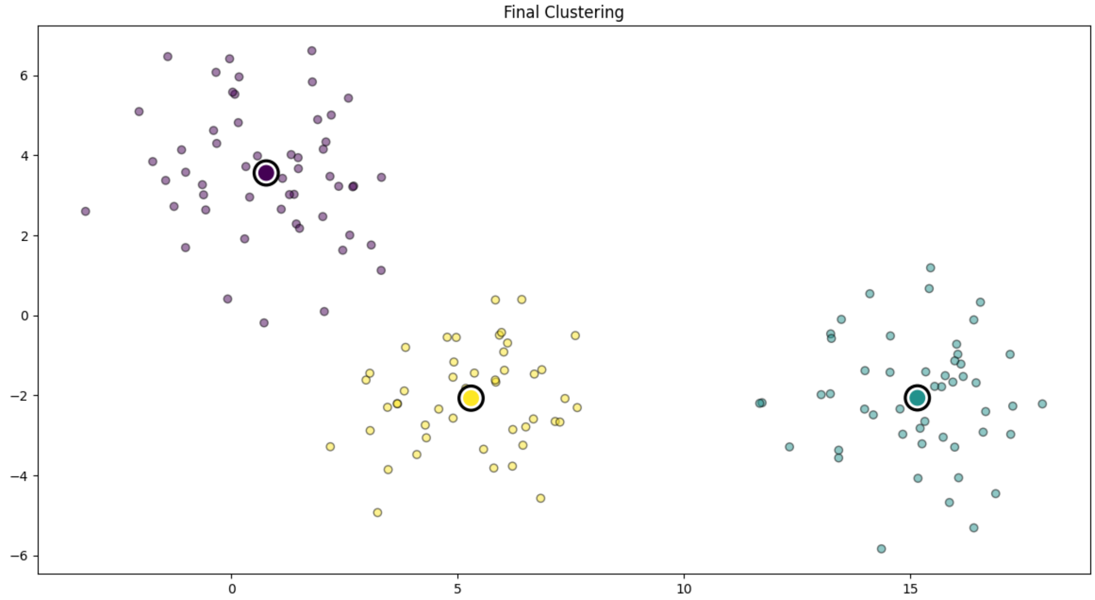
5: Error vs k

As the number of centroids grows larger, the SSE drops significantly. This is because the points present in each cluster are much closer to the centroid than they would be otherwise. For instance, a trial with 150 centroids and 150 points will cluster each point in its own cluster and will have an SSE of 0. Choosing centroid count because of a very low SSE could result in so many clusters that the real patterns in the data are not apparent.



SSE plot for k = 1 to 150, 20 iterations per trial

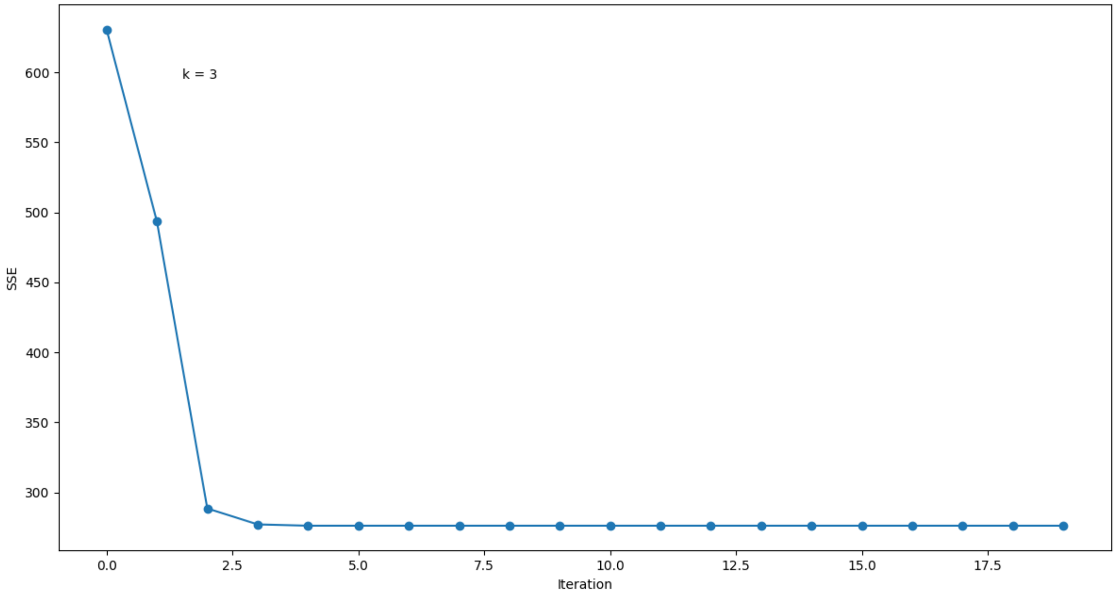
6: Smart initialization



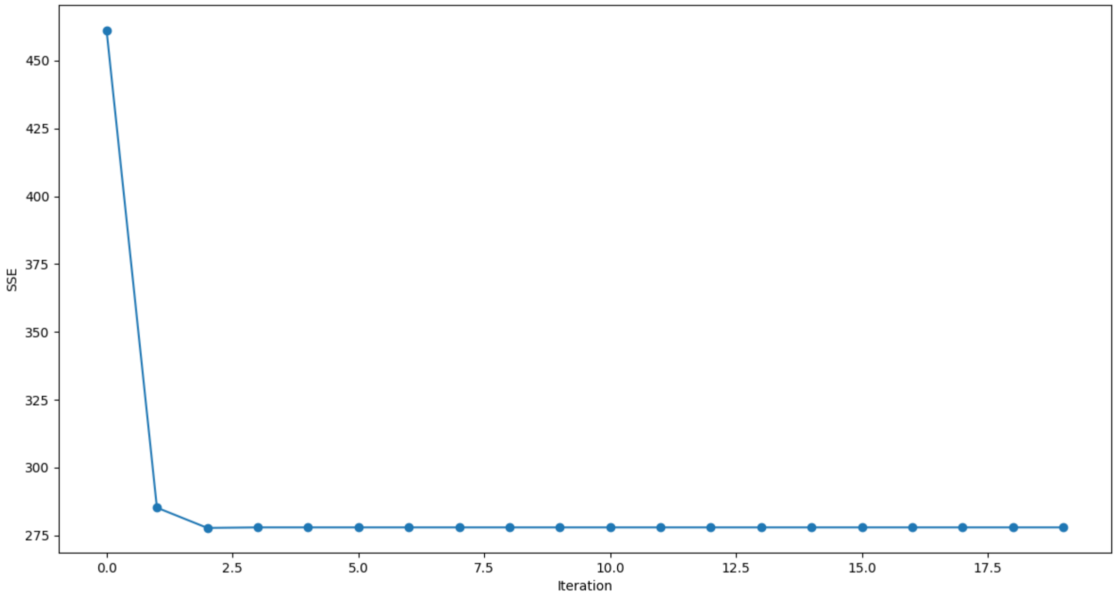
Smart initialization with k = 3

7: Smart initialization vs SSE

I noticed a slight improvement in SSE when using smart initialization. SSE starts at a lower value and decreases much more quickly. However, it seems that both converge to the same SSE value after 4 iterations.



SSE plot for original centroid intialization



SSE plot for smart centroid initialization

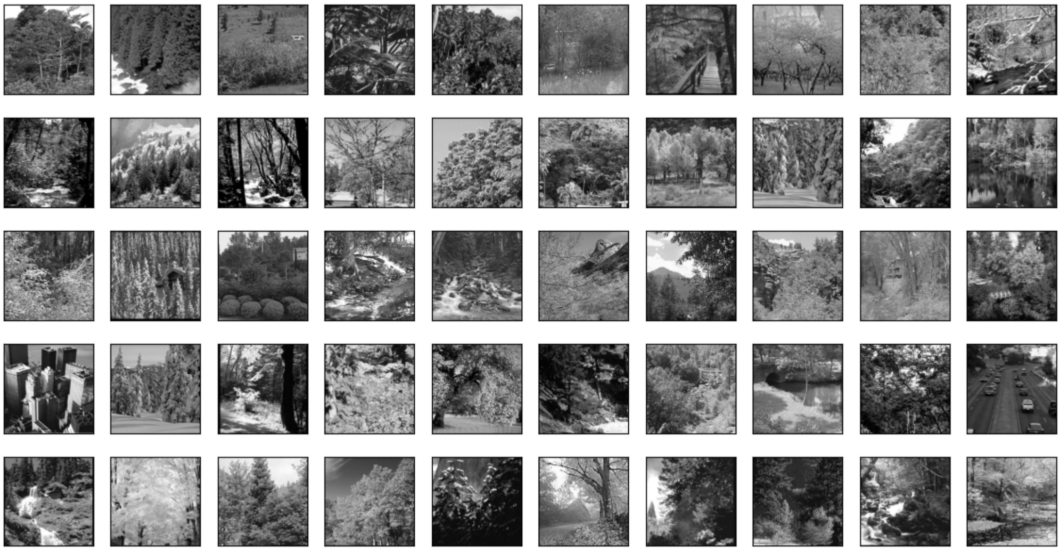
8: Clustering images

a. I used smart centroid intialization.

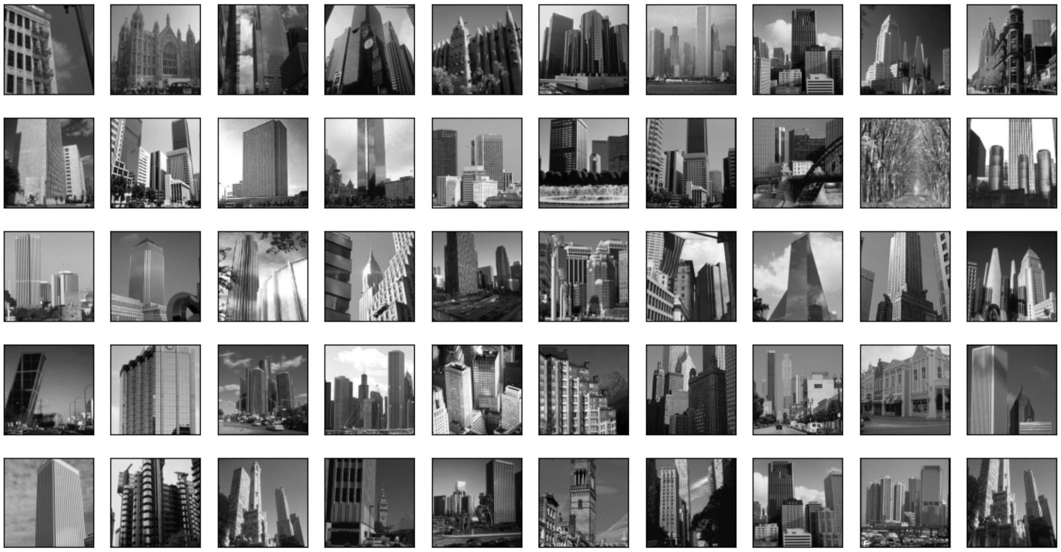
b. The distinct photo types I see appear to be freeways, skyscrapers, and forests. There are more small distinctions between these, such as forests with and without snow, but I think k = 3 would be a good value to try for clustering.

c. I had the best results with k = 4. I still feel that the images are best split into three categories, but there is too much overlap between categories when only three centroids are used.

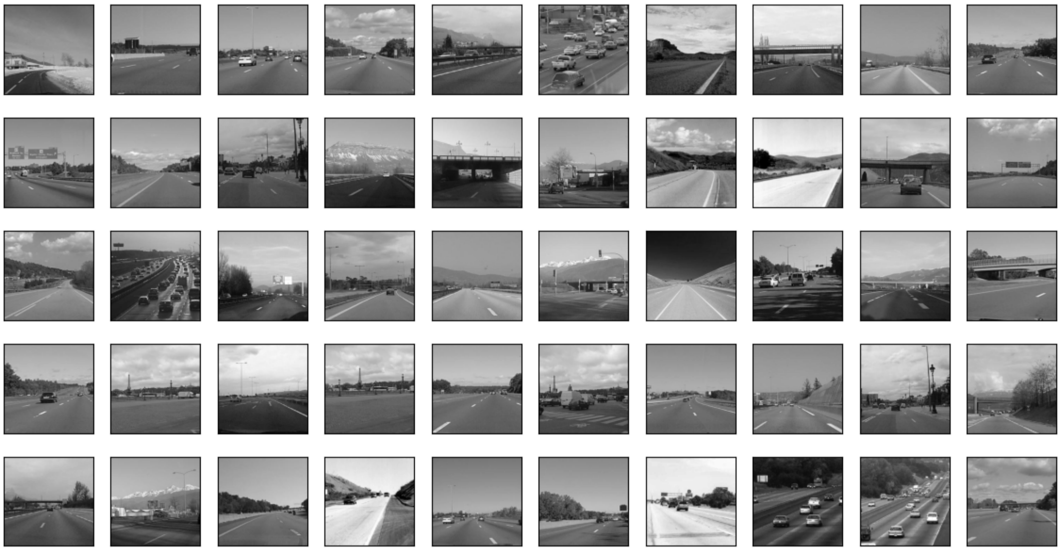
Cluster 1



Cluster 2

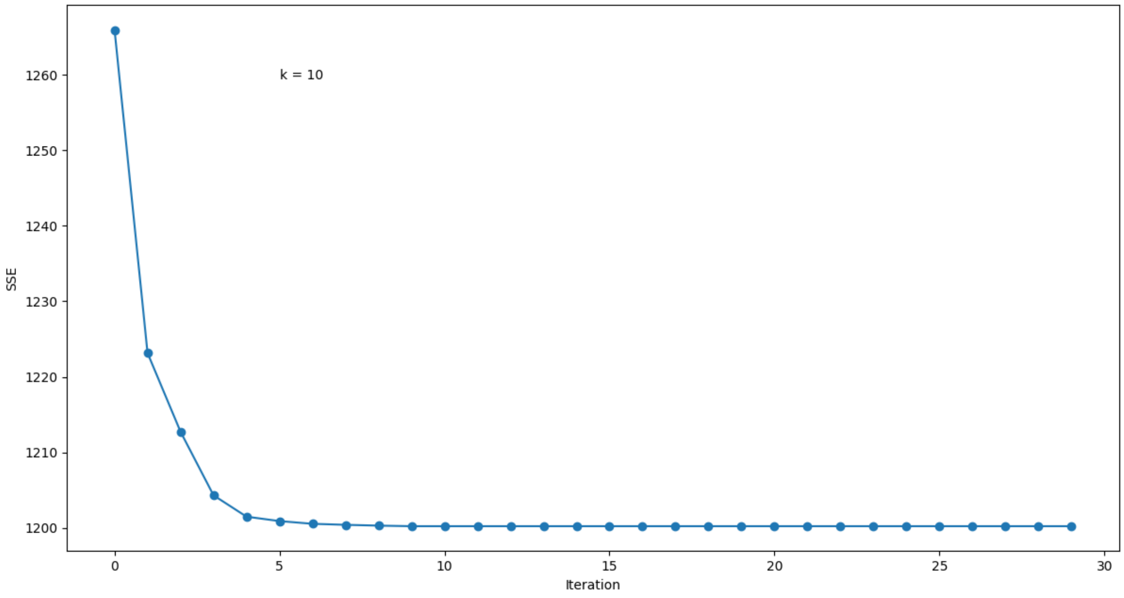


Cluster 3

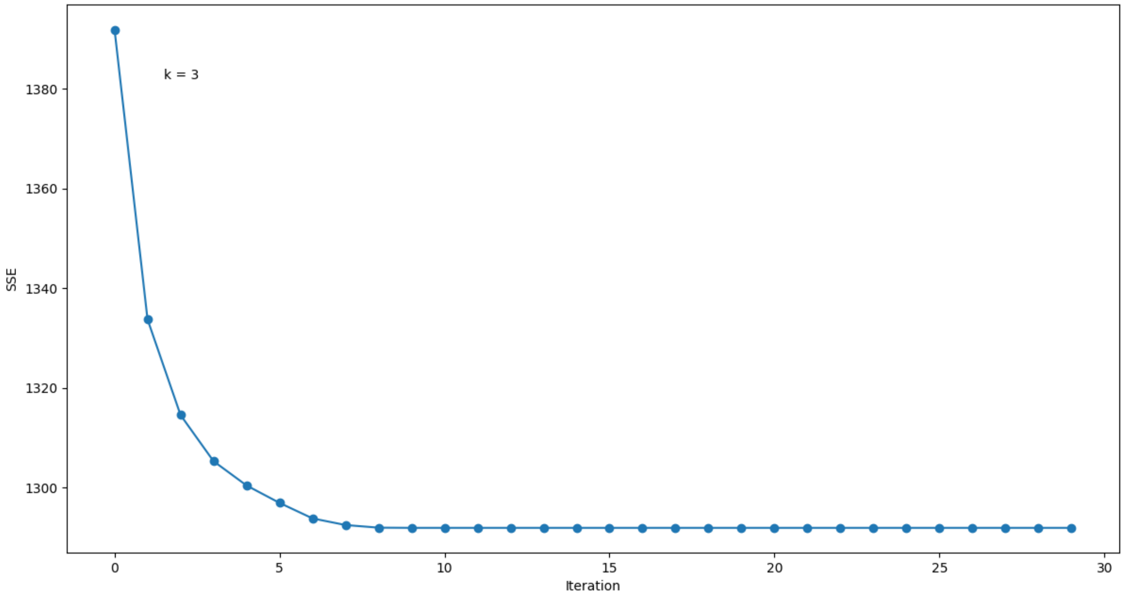


Cluster 4

d. The SSE graph indicates that k = 10 produces more accurate clusters, but this is partially because every point is closer to its respective centroid. In this case, SSE is not a good indicator for clustering quality.



SSE plot for k = 10



SSE plot for k = 3

9: Evaluating clustering as classification
a. I would label the clusters (pictured in 8c) with the following:
Cluster 1: Skyscrapers/cityscapes
Cluster 2: Forests
Cluster 3: Skyscrapers/cityscapes/other buildings
Cluster 4: Roads

b. Impurity: 8%, 4%, 2%, 0%, respectively

10: Debriefing

1. I spent around 10 hours on this assignment.
2. I would rate it as relatively easy, but a little time-consuming.
3. I worked on it alone.
4. I feel I understand most of the material well. The math and concepts are pretty straightforward, especially compared to some other topics covered so far. I'm not certain I did the manual decision tree process correctly but I think I understand the concept well.