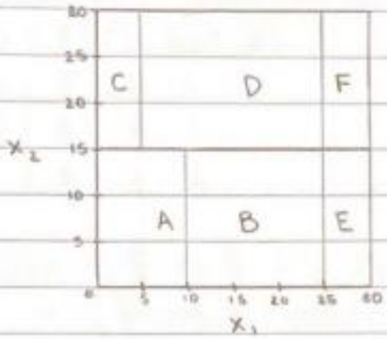
Homework 4 Gretel Rajamoney

Q1:

a)



A belongs to $x_1 \in [0,10)$ $x_2 \in [0,15)$

B belongs to $x_1 \in [10,25)$ $x_2 \in [0,15)$

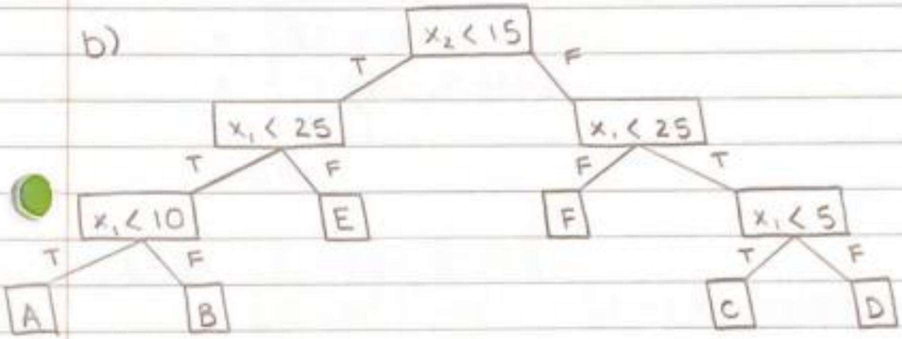
C belongs to $x_1 \in [0,5)$ $x_2 \in [15,30]$

D belongs to $x_1 \in [5,25)$ $x_2 \in [15,30]$

E belongs to $x_1 \in [25,30]$ $x_2 \in [0,15)$

F belongs to $x_1 \in [25,30]$ $x_2 \in [15,30]$

b)



c) It is very difficult to find an accurate decision tree that can perform well on the test data. In order to increase accuracy, reducing the size of the decision tree through removing sections that don't have any power will aid in improving accuracy. By pruning all of the decision nodes, we can reduce overfitting the training data.

Q2: Information Gain for A:

$IG = E(T) - E(T, A)$

A	T	0	1
0	3	2	1
1	3	1	2

$$
\begin{aligned}
E(0) &= -\left[\tfrac{2}{3}\log\tfrac{2}{3} + \tfrac{1}{3}\log\tfrac{1}{3}\right] \\
&= -\left[\tfrac{2}{3}(\log(2)-\log(3)) + \tfrac{1}{3}(\log(1)-\log(3))\right] \\
&= -\left[\tfrac{2}{3}(1-1.58) + \tfrac{1}{3}(0-1.58)\right] \\
&= -\left[\tfrac{2}{3}(-0.58) + \tfrac{1}{3}(-1.58)\right] \\
&= -[-0.386 - 0.526] \\
&= -[-0.9] \\
&= 0.9
\end{aligned}
$$

$$
\begin{aligned}
E(1) &= -\left[\tfrac{1}{3}\log\tfrac{1}{3} + \tfrac{2}{3}\log\tfrac{2}{3}\right] \\
&= -\left[\tfrac{1}{3}(\log(1)-\log(3)) + \tfrac{2}{3}(\log(2)-\log(3))\right] \\
&= -\left[\tfrac{1}{3}(0-1.58) + \tfrac{2}{3}(1-1.58)\right] \\
&= -\left[\tfrac{1}{3}(-1.58) + \tfrac{2}{3}(-0.58)\right] \\
&= -[-0.526 - 0.386] \\
&= -[-0.9] \\
&= 0.9
\end{aligned}
$$

$$
\begin{aligned}
E(Y, A) &= P(0) \cdot E(0) + P(1)E(1) \\
&= \tfrac{3}{6}(0.9) + \tfrac{3}{6}(0.9) \\
&= 0.45 + 0.45 \\
&= 0.9
\end{aligned}
$$

$$
\begin{aligned}
IG &= E(T) - E(T, A) \\
&= 1 - 0.9 \\
&= 0.1
\end{aligned}
$$

Information Gain for B:

$IG = E(T) - E(T, B)$

B	T	0	1
0	2	1	1
1	4	2	2

$$\begin{aligned} E(0) &= -\left(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{2}\log\tfrac{1}{2}\right) \\ &= -\left[\tfrac{1}{2}(\log(1) - \log(2)) + \tfrac{1}{2}(\log(1) - \log(2))\right] \\ &= -\left[\tfrac{1}{2}(0-1) + \tfrac{1}{2}(0-1)\right] \\ &= -\left(-\tfrac{1}{2} - \tfrac{1}{2}\right) \\ &= 1 \end{aligned}$$

$$\begin{aligned} E(1) &= -\left(\tfrac{2}{4}\log\tfrac{2}{4} + \tfrac{2}{4}\log\tfrac{2}{4}\right) \\ &= -\left[\tfrac{2}{4}(\log(2) - \log(4)) + \tfrac{2}{4}(\log(2) - \log(4))\right] \\ &= -\left[\tfrac{2}{4}(-1) + \tfrac{2}{4}(-1)\right] \\ &= -\left(-\tfrac{1}{2} - \tfrac{1}{2}\right) \\ &= 1 \end{aligned}$$

$$\begin{aligned} E(Y, B) &= \tfrac{2}{6}(1) + \tfrac{4}{6}(1) \\ &= \tfrac{2}{6} + \tfrac{4}{6} \\ &= 1 \end{aligned}$$

$$\begin{aligned} IG &= E(Y) - E(Y, B) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Information Gain for C:

$IG = E(T) - E(T, C)$

C	T	0	1
0	3	1	2
1	3	2	1

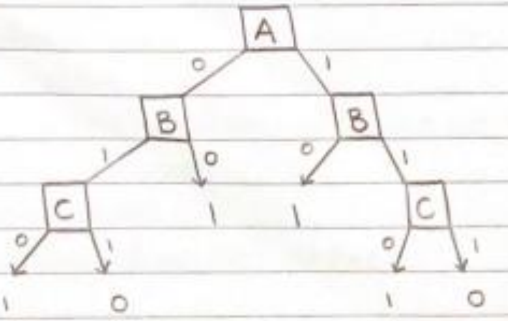
$$\begin{aligned}
E(0) &= -\left(\tfrac{1}{3}\log\tfrac{1}{3} + \tfrac{2}{3}\log\tfrac{2}{3}\right)\\
&= \left[\tfrac{1}{3}(\log(1) - \log(3)) + \tfrac{2}{3}(\log(2) - \log(3))\right]\\
&= -\left[\tfrac{1}{3}(0 - 1.58) + \tfrac{2}{3}(1 - 1.58)\right]\\
&= -\left[\tfrac{1}{3}(-1.58) + \tfrac{2}{3}(-0.58)\right]\\
&= -(-0.52 - 0.38)\\
&= 0.9
\end{aligned}$$

$$\begin{aligned}
E(1) &= -\left(\tfrac{2}{3}\log\tfrac{2}{3} + \tfrac{1}{3}\log\tfrac{1}{3}\right)\\
&= -\left[\tfrac{2}{3}(\log(2) - \log(3)) + \tfrac{1}{3}(\log(1) - \log(3))\right]\\
&= -\left[\tfrac{2}{3}(1 - 1.58) + \tfrac{1}{3}(0 - 1.58)\right]\\
&= -(-0.38 - 0.52)\\
&= 0.9
\end{aligned}$$

$$\begin{aligned}
E(Y, C) &= P(0) \cdot E(0) + P(1) \cdot E(1)\\
&= \tfrac{3}{6}(0.9) + \tfrac{3}{6}(0.9)\\
&= 0.45 + 0.45\\
&= 0.9
\end{aligned}$$

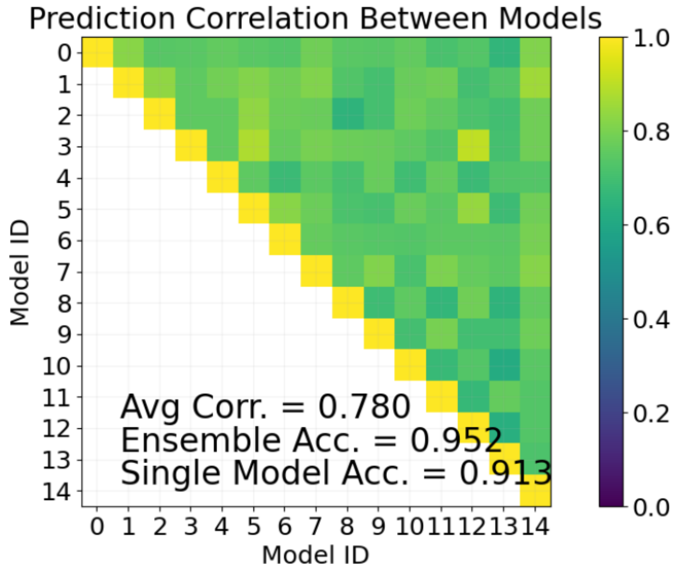
$$\begin{aligned}
IG &= E(T) - E(T, C)\\
&= 1 - 0.9\\
&= 0.1
\end{aligned}$$

The information gain for both candidate splits A and C are equivalent and larger than B. Therefore we can select either of them as the root node.



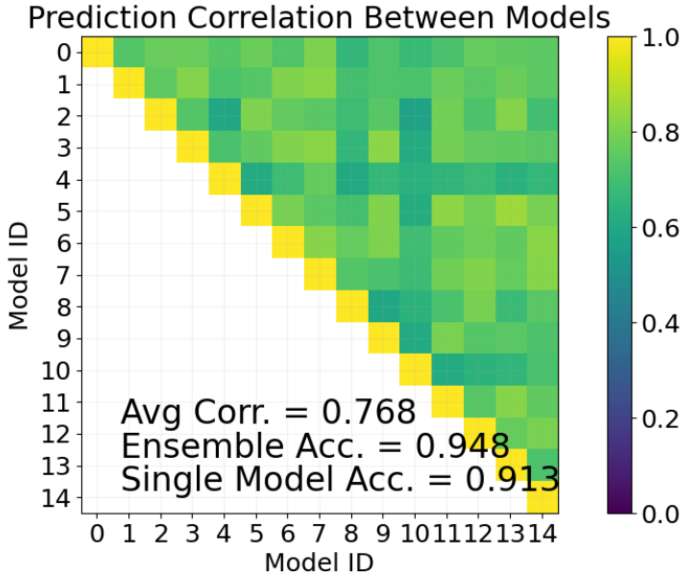
since all Y's can be predicted correctly using our final tree, the accuracy is 100%.

Question 3:



a.

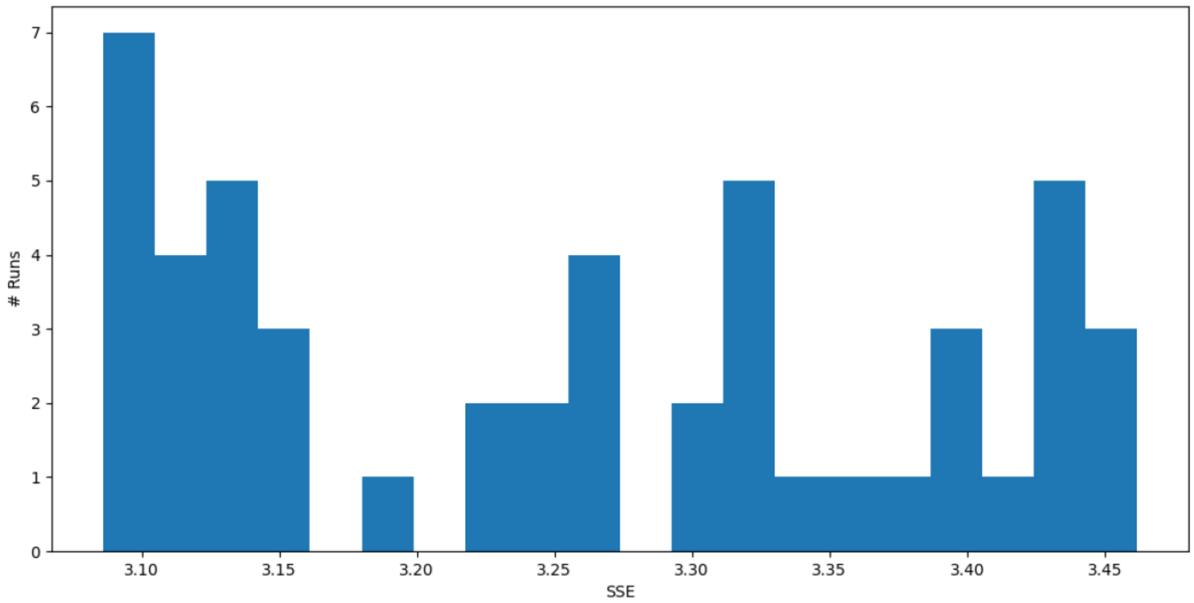
The average correlation in my model dropped from a 0.984 to a 0.780. This is a good thing because uncorrelated errors lead to better ensembles.



b.

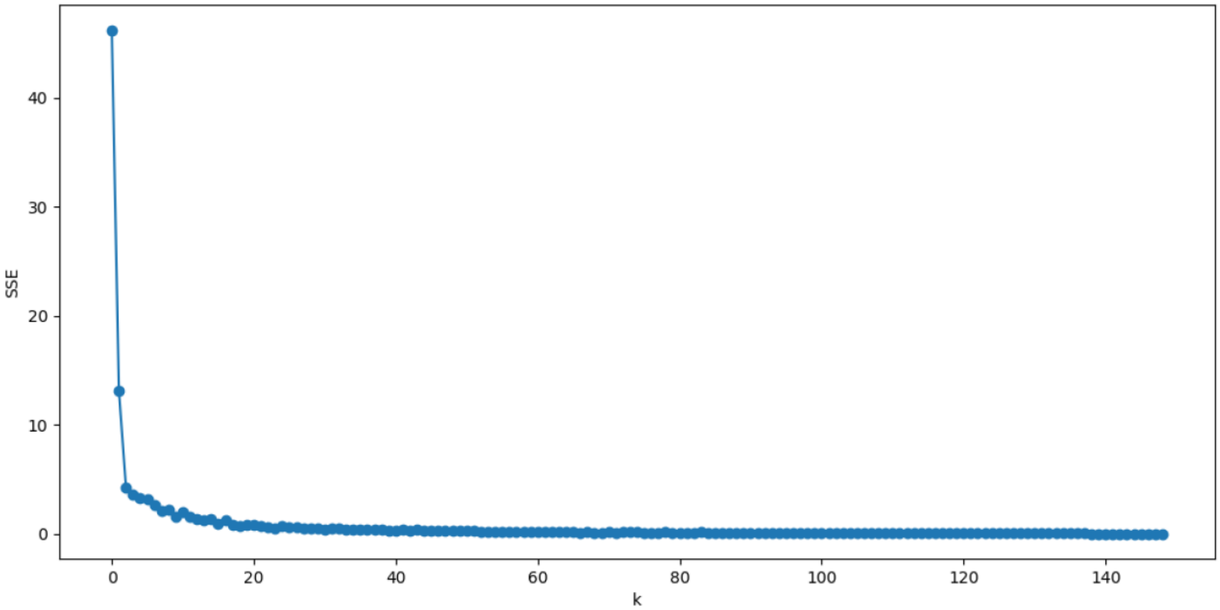
The average correlation in my model dropped from a 0.984 to a 0.768. This is a good thing because uncorrelated errors lead to better ensembles.

Question 5:



The SSE represents the sum of the squared distance between the centroid to each member included within the cluster. Since the SSE remains within the range of 3.10 to 3.45, with a random distribution throughout all of the runs, it is difficult to notice any patterns. This essentially means that the variation between the SSE and all of the programs runs is completely randomized. Since we do want as low of an SSE as possibly, we can analyze this graph in order to select the k values that contain the smallest SSE.

Question 6:

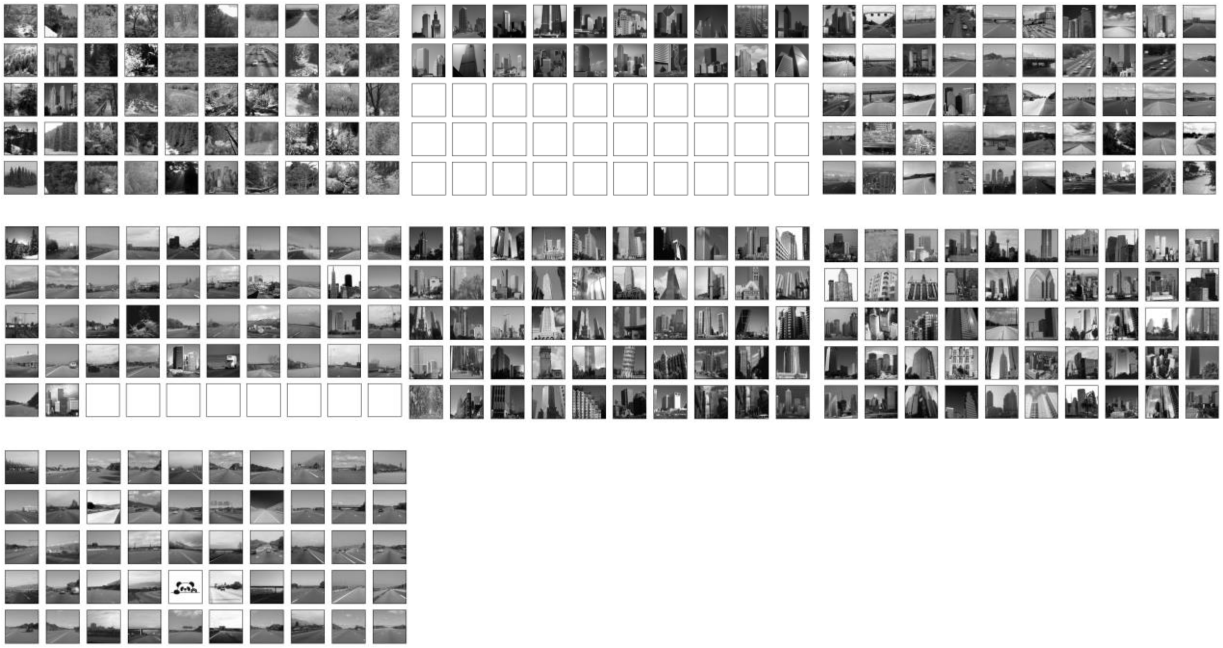


Selecting a k value within k-means clustering is incredibly difficult, but there are techniques that help you select a generally accepted better k value. We can see that as the value of k increases, the sum of squared error decreases. Through using the elbow method, we can select a reasonably optimal value for k by choosing the value where the k abruptly changes from rapidly decreasing to essentially arriving at a asymptote. Using the graph above as an example of this technique, the approximate k value that would generate the best results is 7.

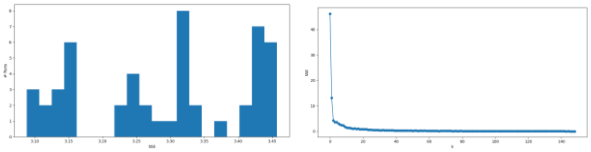
Question 7:

a. There are a few types of images displayed, they include: skyscraper buildings, trees, and highways. The images displayed in the datasets appear to be fine, and k=10 is a reasonable parameter for the dataset. But the dataset can be improved if the default parameter of k=10 is reduced to a slightly smaller value.

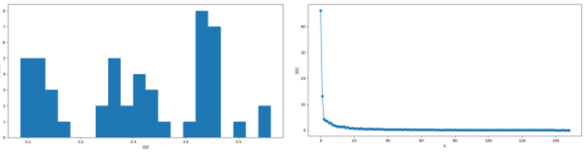
b. The default hyperparameter of k=10 was reasonable for the dataset, but in order to explore other possibilities, I set the parameter to the value of k=3. Although the outputted images were still acceptable, there were a lot more misassigned images within each cluster. For example, clusters that were mainly skyscrapers, contained a lot more trees and highways when the k was set to 3 as opposed to when it was set to the default value of 10.



c. K = 10



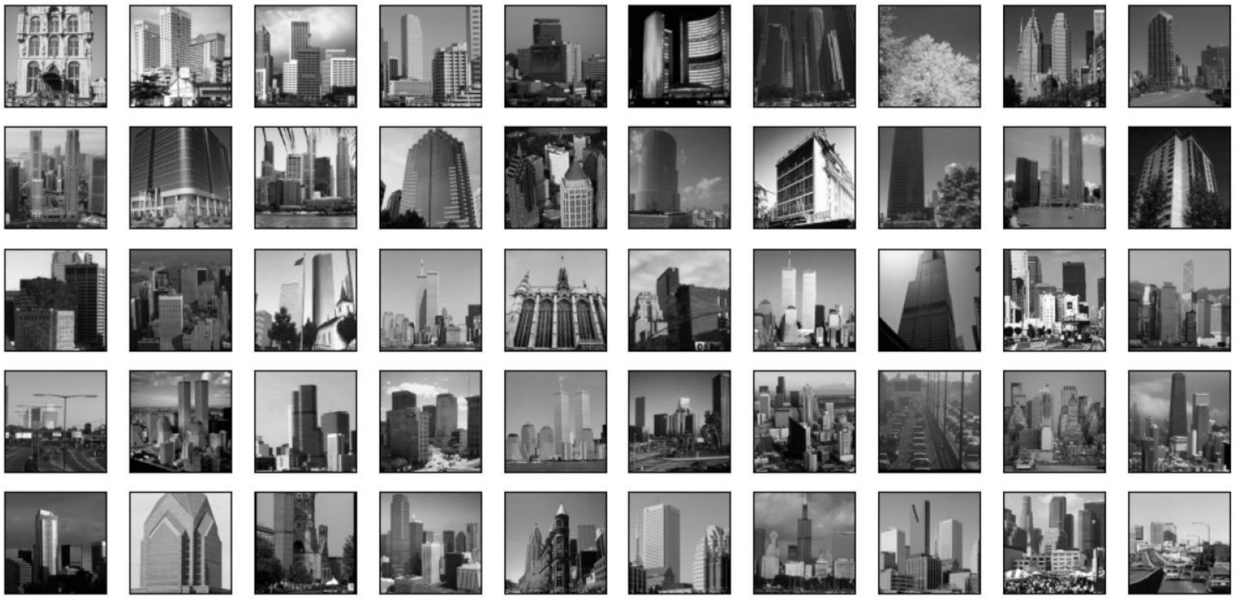
K = 7



Sum of squared error is not a good indicator of clustering quality, this is why selecting a k value through analyzing the SSE graph is not ideal. Even though the quality of the clusters were significantly better when the k hyperparameter was changed from the default value of 10 to a new value of 7, the sum of squared error graphs indicate otherwise.

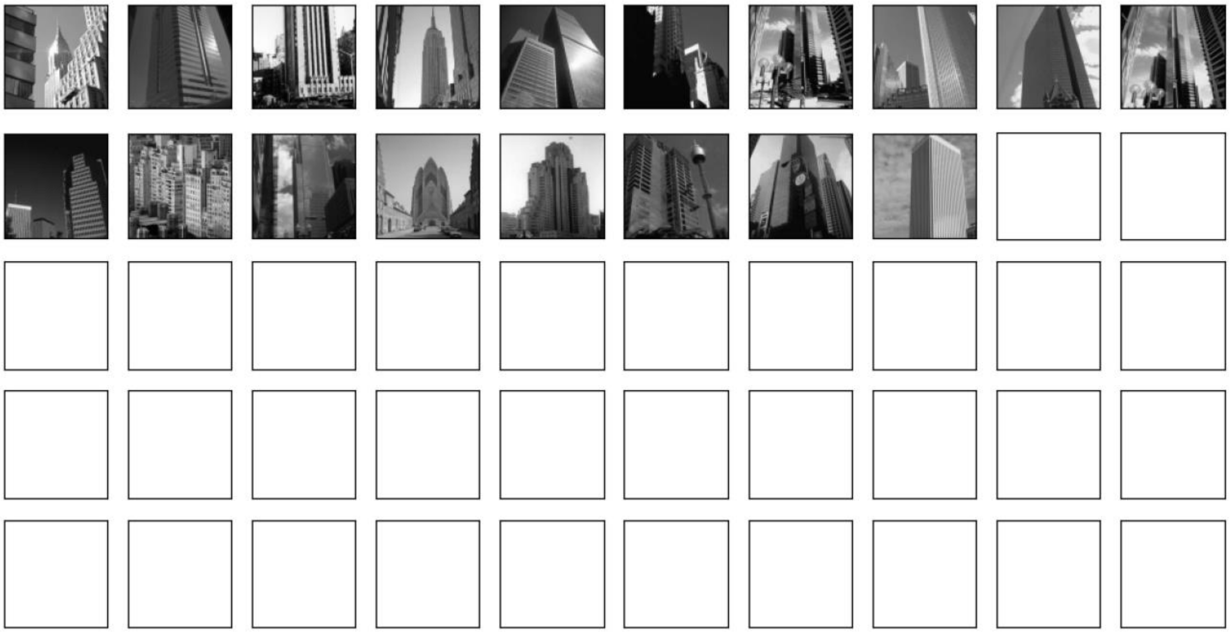
Question 8:

Label 1: Skyscrapers



Purity: 4/50

Label 2: Skyscrapers



Purity: 0/18

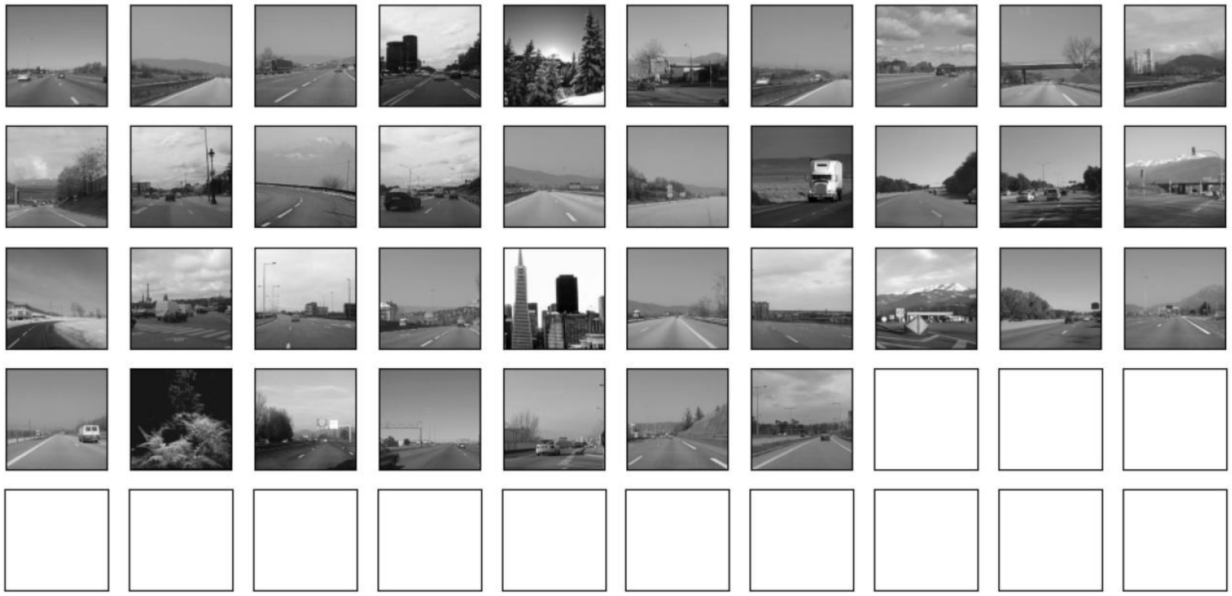
Label 3: Skyscrapers

Purity: 0/11

Label 4: Trees

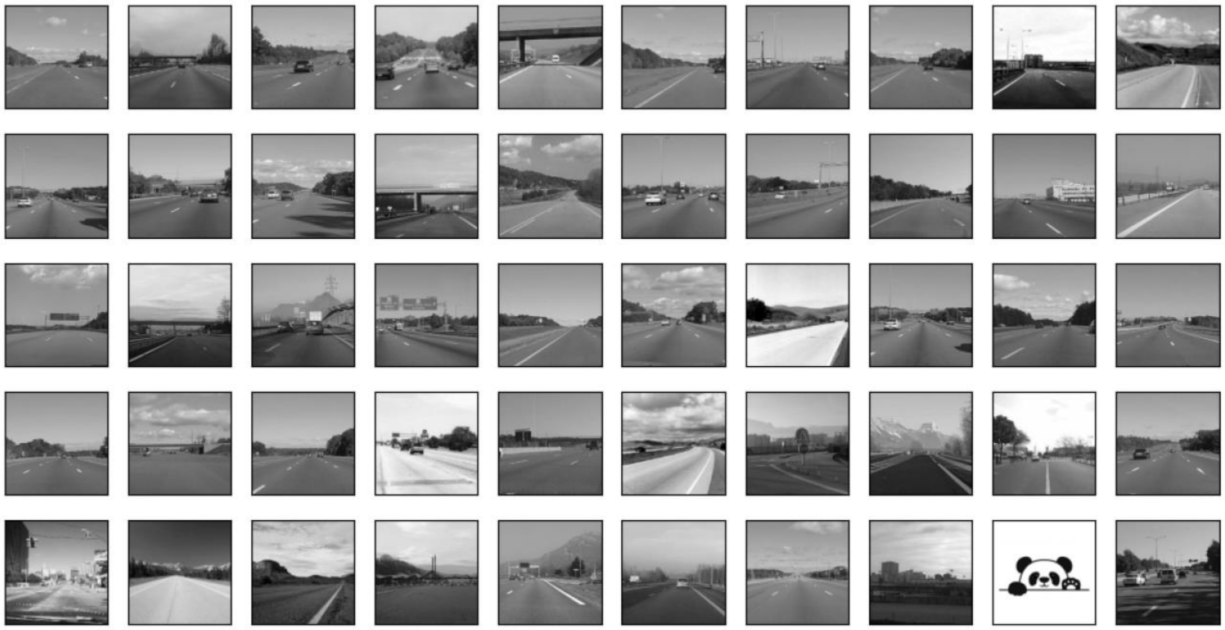
Purity: 2/50

Label 5: Highways



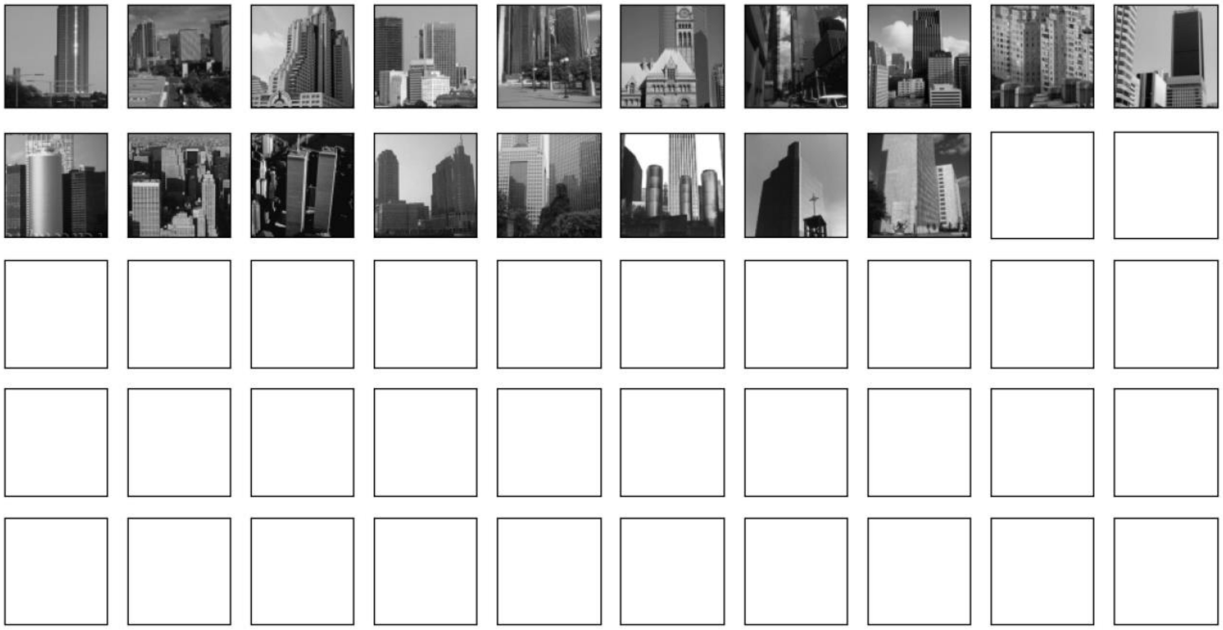
Purity: 3/37

Label 6: Highways



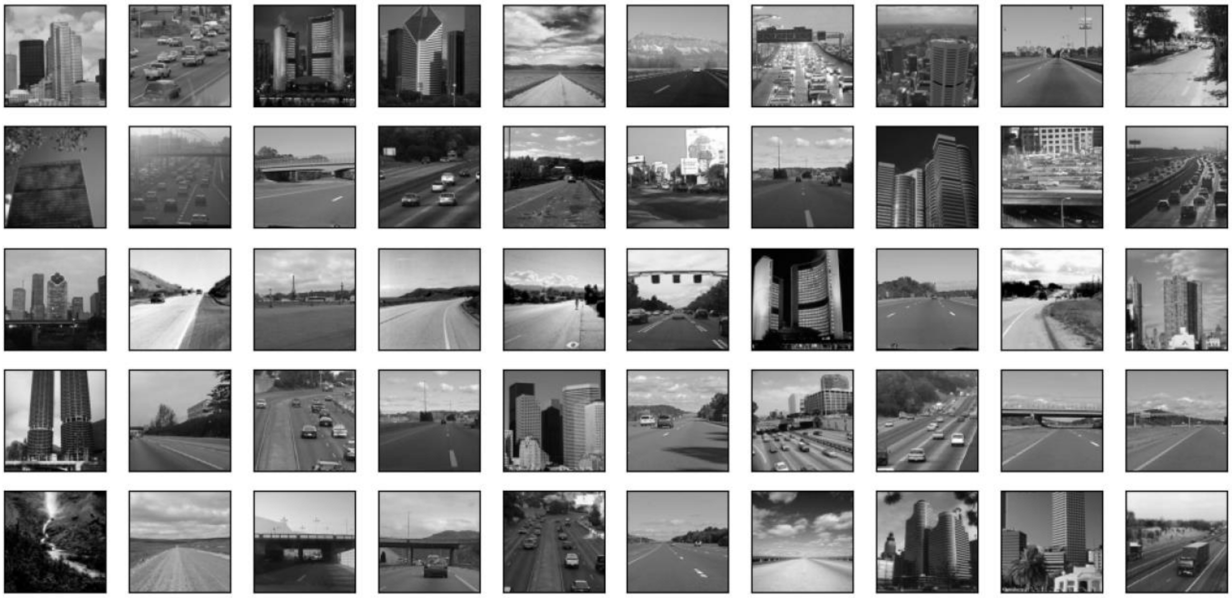
Purity: 1/50

Label 7: Skyscrapers



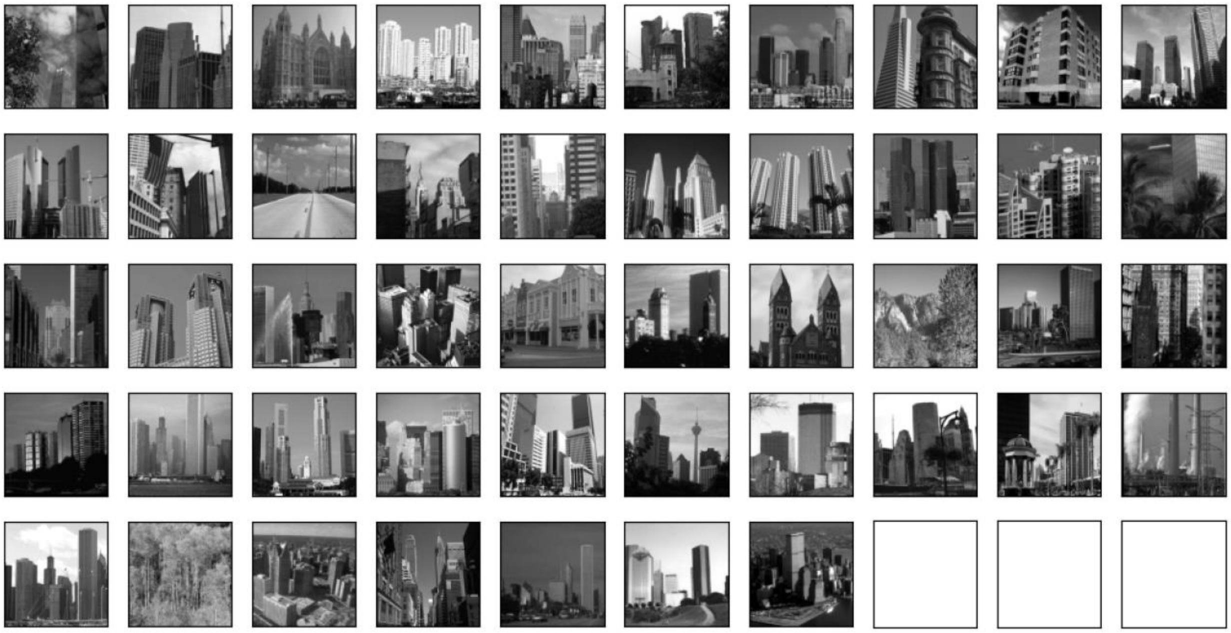
Purity: 0/18

Label 8: Highways



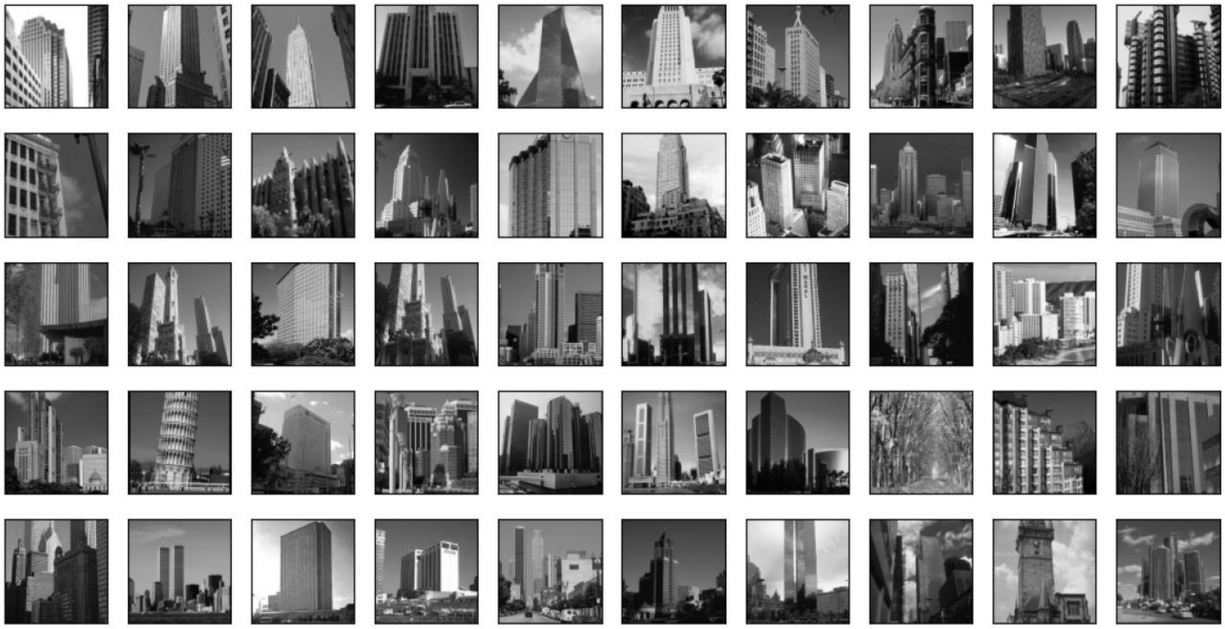
Purity: 15/50

Label 9: Skyscrapers



Purity: 5/47

Label 10: Skyscrapers



Purity: 1/50

Debriefing:

1. I spent approximately 12 hours on this assignment.
2. I would rate it moderate.
3. I worked mainly alone but discussed a few things with a peer in order to confirm that the outputs were reasonable as well as comparable to theirs.
4. 65%
5. None