Part 1:

Results:

Random initialization. Choose the initial centers randomly.

From left to right: original image, k-means compressed images with k = 2, 5, 10, 15, 20.

File sizes (kb): top: 763, 122, 177, 159, 163, 158; bottom: 760, 69, 107, 115, 99, 122

Average compression ratio and standard deviation: top: 4.973, 0.7404; bottom: 7.727, 1.917

Ordered slice initialization. Choose the initial centers by ordering the pixel values and partitioning the array into k slices and taking a pixel value at the start of each slice.

File sizes (kb): top: 763, 126, 171, 162, 157, 154 ; bottom: 760, 83, 105, 119, 114, 119

Average compression ratio and standard deviation: top: 5.006, 0.615; bottom: 7.167, 1.165

Averaged initialization. A more sophisticated approach may be to use average pixel values when partitioning the array.

1. order the pixel values
2. begin iterating over the pixel values, keeping a running average
3. if a pixel value is reached that is sufficiently different from the average, set one of the initial centers as this average and begin a new running average starting at this pixel value
4. repeat steps 2 & 3 until you have k initial centers

File sizes (kb): top: 763, 127, 153, 162, 158, 160; bottom: 760, 89, 110, 127, 117, 125

Average compression ratio and standard deviation: top: 5.065, 0.5371; bottom: 6.802, 1.038

Discussion:

There is a tradeoff between image quality and compression, naturally. As there are fewer clusters, there is less information in the image. The file size does decrease with fewer clusters, but not by very much (about 80kb with k = 2 vs 120kb with k = 20, k-values higher than 2 seem to lean toward the same compression ratio as k = 20). It's clear to me what the images are even with k = 2, so choosing a k is highly dependent on what the aim of the compression is. A value of k = 15 seems to give an image of sufficient photographic quality, and I would say that it gave the best results in each method. The averaging method for choosing initial clusters gives better results when there is more variance in the image, while the slicing method seems to give better results when there is less variance.