CS 4375
ASSIGNMENT 3

Name of students in your group:

James Hooper (jah171230)
Hritik Panchasar (hhp160130)

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

For Part 1:

https://en.wikipedia.org/wiki/Elbow_method_(clustering)
https://en.wikipedia.org/wiki/Silhouette_(clustering)
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

For Part 2:

https://en.wikipedia.org/wiki/Color_quantization
https://lmcaraig.com/color-quantization-using-k-means/
https://appliedmachinelearning.blog/2017/03/08/image-compression-using-k-means-clustering/

James Hooper
Hritik Panchasara

Assignment 3 Report

Part 1:

SSE & k-value Table

Experiment Number	Value of k	SSE Value
1	1	143183.99999999965
2	2	92214.38198986319
3	3	68072.02731593423
4	4	53123.002362695406
5	5	44219.74030806203
6	6	39137.954965978504
7	7	34352.30367095612
8	8	31194.121180412967
9	9	29189.814134817414
10	10	27481.080433025574

- According to the Elbow Method, where one chooses the proper k-value by plotting the value of k versus the SSE Value then analyzing which k value causes the 'elbow' joint, the proper value to be used is a k-value of 2. We decided to also attempt the Silhouette Score method, where one chooses the proper k-value by plotting the value of k versus the Silhouette Score then choosing the which k-value gives the highest score, which implied that the k-value of 2 was the best choice. The reason we even got to the point of implementing another way to discern the proper value of k was due to not being fully satisfied with the results from the Elbow Method. It was very difficult to determine which value of k was best based solely off the Elbow Method graph. This could maybe be solved by better preprocessing. A side note to be mentioned is that using different scaling such as Normalization, Standardization, and the Min-Max Scaler gave similar graphical trends as even just using the raw data.
- **Language/Tools/Packages:**
    - *Language:* Python
    - *Tools/Packages:* Pandas, Sklearn, & Matplotlib
        - sklearn.cluster, sklearn.metrics, preprocessing from sklearn
        - matplotlib.pyplot

James Hooper
Hritik Panchasara

Part 2:

Report Table

Image	k-value	Image Quality	Original Size	Quantized Size	Time to Quantize
1	4	Very blurred. Weird color effect.	296 KB	285 KB	20.466s
2	4	Blurred and almost mono-coloured.	407 KB	394 KB	23.727s
3	4	Low Quality reimage. Very blurred.	189 KB	189 KB	13.710s
1	8	Cartoonish of original.	296 KB	297 KB	51.946s
2	8	More similar to the original but still very choppy.	407 KB	394 KB	59.285s
3	8	Blurred out version of original	189 KB	202 KB	45.707s
1	16	Closer to original but the whites still blurred.	296 KB	292 KB	114.594s
2	16	Background and scenery look way better, but the water is blurred.	407 KB	397 KB	132.724s
3	16	Overall cartoonish & blurred. Main player of focus looks pretty good	189 KB	220 KB	92.742s

- The results of this code came out as expected with the larger k-values making an image closer to the original. The size increasing as the k-value increased does make sense, but it doesn't make complete sense that it would have a bigger file size than the original. An interesting note is that for k=8 all of the new sizes increased from the original for image 1 & 3, but when the k-value was increased to 16 the file size for image 1 actually went down. These oddities may be due to the selection and random state used. Mentioning that, it should be stated the random_state stayed at a constant value of 0.
- The "best" value of k seems to be 4 if one wanted an okay image for less size, but the size difference for different k-values seems negligible from a general perspective. According to the observed results, certain k-value sizes work best for different images for quality & filesize.
- **Language/Tools/Packages:**
    - *Language:* Python
    - *Tools/Packages:* Numpy, Sklearn, Skimage, Time, & Matplotlib
        - cluster from sklearn
        - io from skimage
        - matplotlib.pyplot
        - time from time