# Data Mining Homework: Clustering - Report

## Topic Selection Journey

When starting this assignment, I faced a key question: What kind of problem is truly suitable for clustering?

I found that many common clustering assignment topics, such as the Iris and Wine datasets, actually have clear labels and are essentially classification problems that should be solved using supervised learning. Using clustering on these topics feels like "forcing clustering for the sake of using clustering."

I explored various possible applications: customer segmentation, hotspot detection, music classification, anomaly detection, etc. However, after deeper consideration, I found that many of these problems share a common issue: there are actually better solutions, or clustering is merely an auxiliary tool rather than the core method.

Ultimately, I chose image compression (K-means Color Quantization) for a simple reason:

1. This is clustering's purest application - K-means was originally designed for vector quantization, and image compression is its classic use case
2. The K value has clear practical meaning - It's not randomly guessing how many clusters to use, but directly corresponds to "how many colors to preserve," which affects compression ratio and visual quality
3. Results can be directly verified - Compression effects are immediately visible, requiring no complex evaluation metrics
4. No "better alternative methods" - For color quantization, clustering is the standard solution

## Result

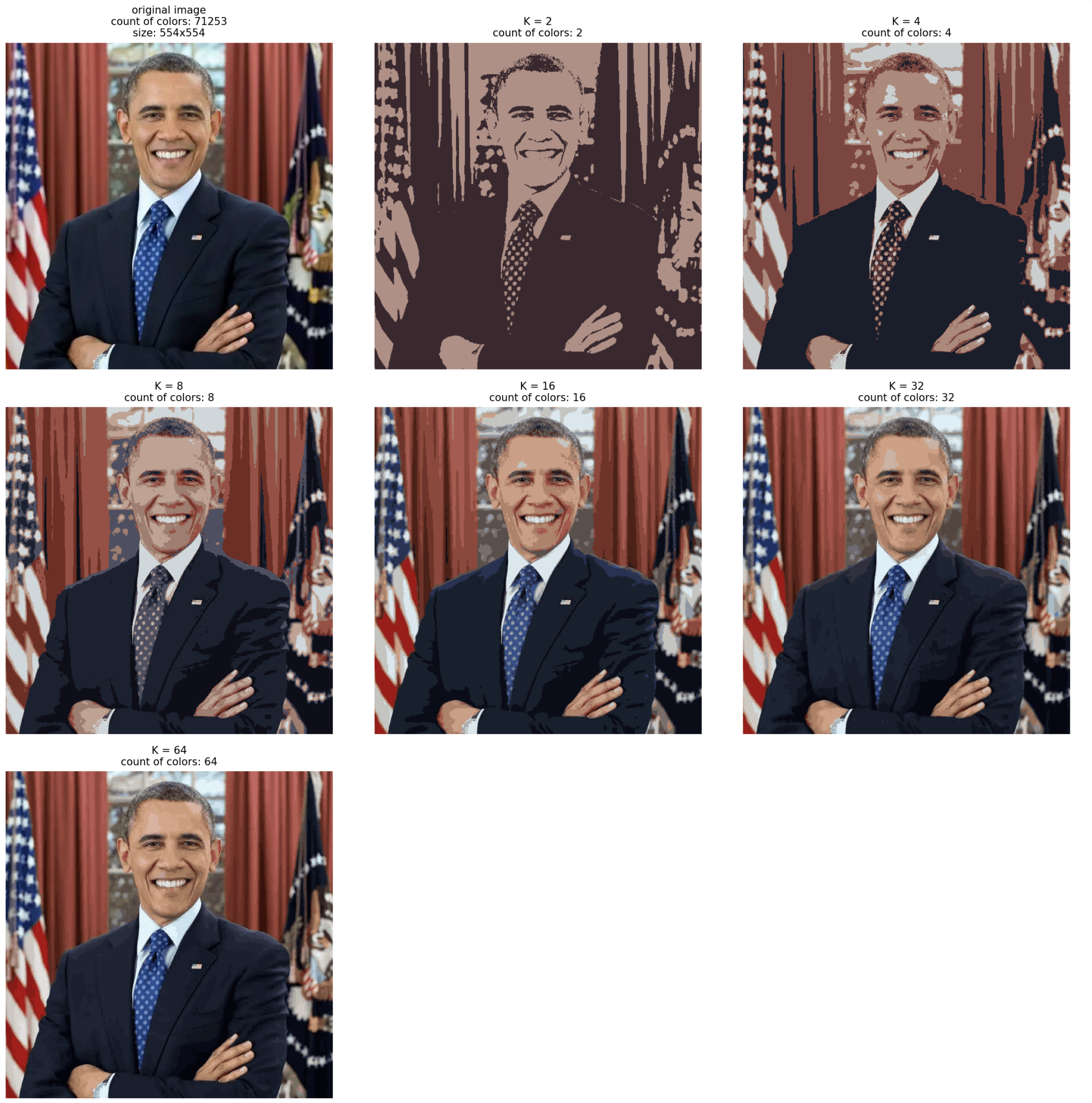
### Experimental Setup

I used a portrait photo of Barack Obama as the test image for K-means color quantization. The original image contains **71,253 unique colors** with dimensions of **554×554 pixels**.

I tested six different K values: 2, 4, 8, 16, 32, and 64, representing different compression levels. Each K value corresponds to the number of colors retained in the compressed image.

### Visual Results

The comparison clearly demonstrates the trade-off between compression ratio and visual quality:



**Key Observations:**

- **K=2**: Extreme compression with only 2 colors creates a high-contrast, binary effect. The image is reduced to basic silhouettes with almost all details lost, creating an artistic poster-like appearance.
- **K=4**: Still heavily compressed, showing severe posterization. Only the most basic color regions (skin, suit, background) are distinguishable, with harsh boundaries between color blocks.
- **K=8**: Color banding is very noticeable, especially in skin tones and the background curtains. The image has a cartoonish appearance but facial features become

recognizable.

- **K=16**: A significant improvement - facial features are clear, the smile is natural, and the overall composition is understandable. However, color transitions are still somewhat abrupt, particularly in gradient areas.
- **K=32**: The image looks quite natural. Color gradients in skin tones are smoother, the suit's texture is more refined, and most viewers would find this acceptable quality for casual viewing.
- **K=64**: At normal viewing distance, this is nearly indistinguishable from the original image. The subtle color variations in skin tones, shadows, background details, and the American flag are all well-preserved.

## Quantitative Analysis

| K Value | Colors Used | Compression Ratio | Visual Quality | Notes |
|---|---|---|---|---|
| 2 | 2 | 35,627:1 | Very Poor | Binary/silhouette effect |
| 4 | 4 | 17,813:1 | Poor | Severe posterization |
| 8 | 8 | 8,907:1 | Fair | Noticeable color banding |
| 16 | 16 | 4,453:1 | Good | Recognizable but simplified |
| 32 | 32 | 2,227:1 | Very Good | Natural-looking |
| 64 | 64 | 1,113:1 | Excellent | Nearly identical to original |

**Compression ratio = Original colors (71,253) / K**

## Finding the "Sweet Spot"

The results reveal an interesting pattern in the quality-compression trade-off:

- **Dramatic improvements from K=2 to K=16**: Each doubling of K brings substantial and immediately noticeable visual improvements. The jump from K=8 to K=16 is particularly significant, transforming the image from "cartoonish" to "acceptable."

- **Diminishing returns beyond K=32**: While K=64 is technically superior, the perceptual difference from K=32 is minimal for most viewing contexts. This demonstrates the logarithmic nature of human color perception.

- **K=64 as the perceptual threshold**: At this point, without zooming in or direct side-by-side comparison, the compressed image appears virtually identical to the original. We've reduced 71,253 colors to just 64 while maintaining excellent visual fidelity - **a compression ratio of over 1,000:1**.

For this particular portrait image, the optimal choice depends on the use case:

- Choose **K=16** for maximum compression when quality is secondary (over 4,000:1 compression)
- Choose **K=32** for the best balance of compression and quality (over 2,000:1 compression)
- Choose **K=64** when visual fidelity is paramount (over 1,000:1 compression)

## Why This Validates the Topic Choice

This experiment perfectly demonstrates why K-means color quantization is an ideal clustering application:

1. **Clear, practical objective**: We're not arbitrarily guessing the "right" number of clusters - we're consciously choosing how many colors to retain based on concrete quality and compression requirements.
2. **Immediate, intuitive verification**: Unlike abstract clustering tasks requiring complex evaluation metrics, we can simply view the images and judge quality. The results are self-explanatory.
3. **Real-world relevance**: This technique is actively used in practical applications:
    - GIF format (limited to 256 colors)
    - Web image optimization
    - Mobile app assets
    - Retro game graphics
    - Reducing memory footprint in embedded systems
4. **Pure clustering problem**: There's no "ground truth" labels we're ignoring. We're using clustering for its original, intended purpose: **vector quantization**. K-means was literally designed for this type of problem.
5. **Demonstrates core clustering concepts**: This application clearly shows key clustering principles:
    - How K affects granularity of grouping
    - The trade-off between model complexity (K) and representation accuracy

The success of this experiment - achieving visually excellent results with just 64 colors out of 71,253 - proves that K-means effectively identifies the most representative color clusters in the image data.