Data-gathering, Stakeholders, & KPIs

Stakeholders: Art curators, people looking for similar art pieces, scholars

KPIs:

- Measure how well the k-means clustering algorithm captures dominant colors in the paintings by calculating the distance between cluster centroids of the input and recommended images.
- Assess how accurately the Gabor filters capture texture patterns by calculating cosine similarity between the Gabor filter responses of the input and recommended images.
- Manually check whether recommendations are as similar to the input as we expect them to be
- Measure the time taken to extract colors and textures/apply ViT and recommend an image.
- Measure whether the system suggests a diverse range of images or if it tends to favor similar patterns repetitively.

Datasets:

- National Gallery of Art:
    - Images available in high resolution (via API calls), but some might be smaller resolution under the fair use doctrine.
    - 118,742 published images total: drawing (27.6%), painting (3.3%), technical material (0.06%), print (41.6%), decorative art (0.7%), portfolio (0.8%), etc.
    - Features: Title, years displayed, medium, dimensions, artist, artist nationality, attribution
        - Theme and type of painting and forward text with historical context is included but largely missing.
- Kaggle:
    - 30GB of paintings scraped from WikiArt and classified into movement/style
    - There can be mislabeling / misclassification of the arts
    - Some style are specific, say Baroque, but some other can be broadened (eg Byzantine -> Western Medieval)
- Huggingface WikiArt:
    - 81,444 images are available. Approximately 51.9% of the images have a width between 1,170 and 1,580 pixels. The dataset includes a wide range of resolutions, predominantly featuring higher-resolution images suitable for detailed analysis.
        - 129 artists including unknown, 11 genres including unknown and 27 different styles.
        - Genres ex 'landscape'', 'cityscape', 'portrait' etc
        - Styles ex 'Abstract_Expressionism', 'Action_painting', 'Baroque' etc
    - Artists only include the name, no years no detailed information.