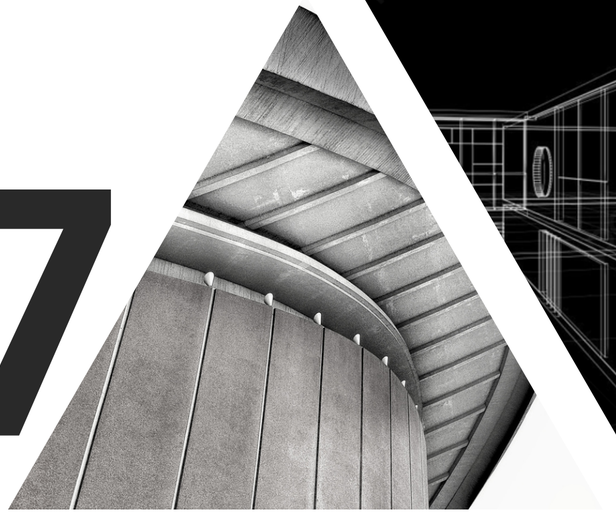
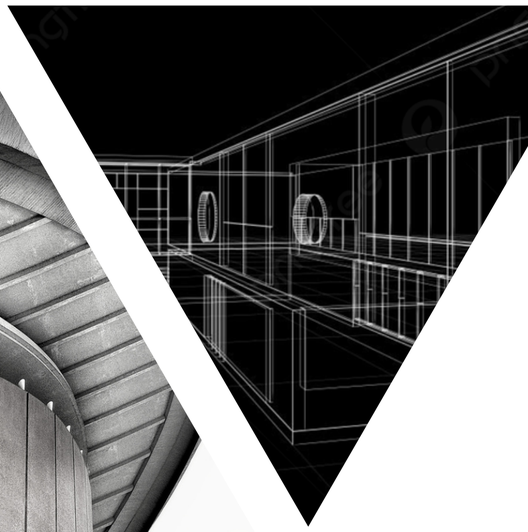
DS 203
Project E7





Group Number : 55
Nidhi Goyal (22B0337)
Padmaja Bodavula (22B2207)

Problem Description

Our goal to digitalize building layout designs
A repository of 1183 layout are provided

Design Family formation

Grouping layouts into families, based on their shapes , which which may help to standardize and create design templates

Complexity Classification

Classifying given layouts based on their complexity into low complexity, medium complexity, high complexity

Layout Retrieval

Creating a model to retrieve relevant prior layouts based on a set of parameters provided by the architect

Solution Approach and Achievements

Analyzing given data and removing duplicate images present in the dataset

Forming tight fitting boxes and extracting relevant features from it

Feature Engineering for further grouping and classification models.

Trying different approaches to group layouts into families based on their shapes

Classification of layouts on the basis on their complexities

Creation of model to retrieve design/layouts based on certain parameters provided

Pre-processing

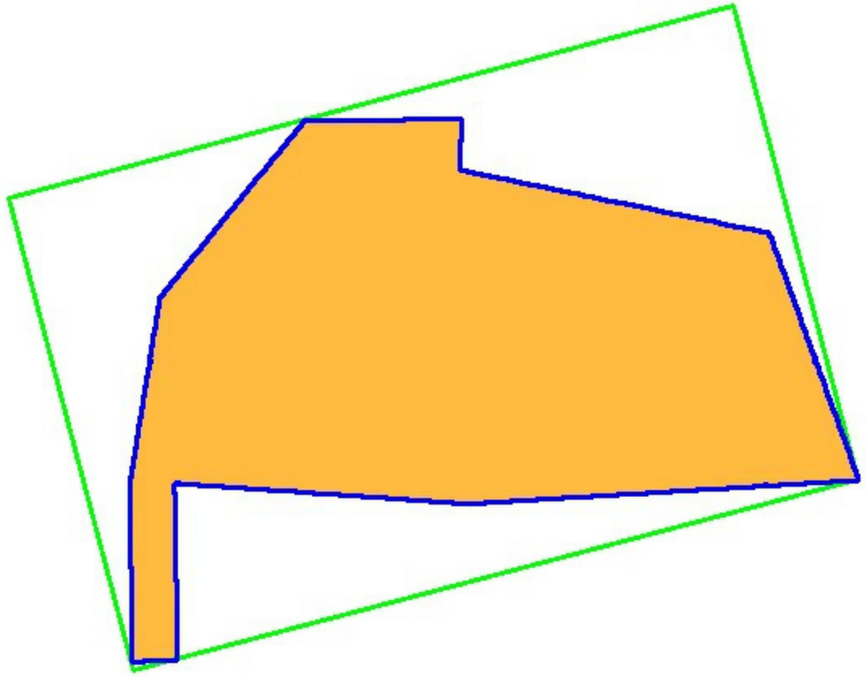
Initially, the dataset provided had 1183 images of building layout designs

We checked for duplicate images present by hashing technique, and removed them

At the end, we were left with 173 unique images of building layout designs

Tight Fitting Box

Sample Output



Inner dark blue polygon
Contour / layout

Green box
The tight-fitting box covering the contour

Features extracted from Tight-fitting boxes

- Dimensions of tight-fitting box
- Area of tight fitting box

Feature Engineering

Contour area

Number of Vertices

Perimeter

Minimum angle change (change in angle between the edges)

Maximum angle changes

Mean angle change

Ratio of area of contour to tight bounding rectangle

Symmetry (calculated using perimeter and average distance from centroid)

Average magnitude of gradient of Laplacian of the image

Design Family formation

Method 1

Principal Component Analysis

Based on the features extracted, we performed PCA analysis to form optimum set of features for classification of images

K-means clustering

Using these features, we performed K-means clustering algorithm to form required number of groups of images

Model Evaluation and Optimal Number Of Clusters

Silhouette Score

It is a measure of how similar an image is to its own cluster compared to other clusters

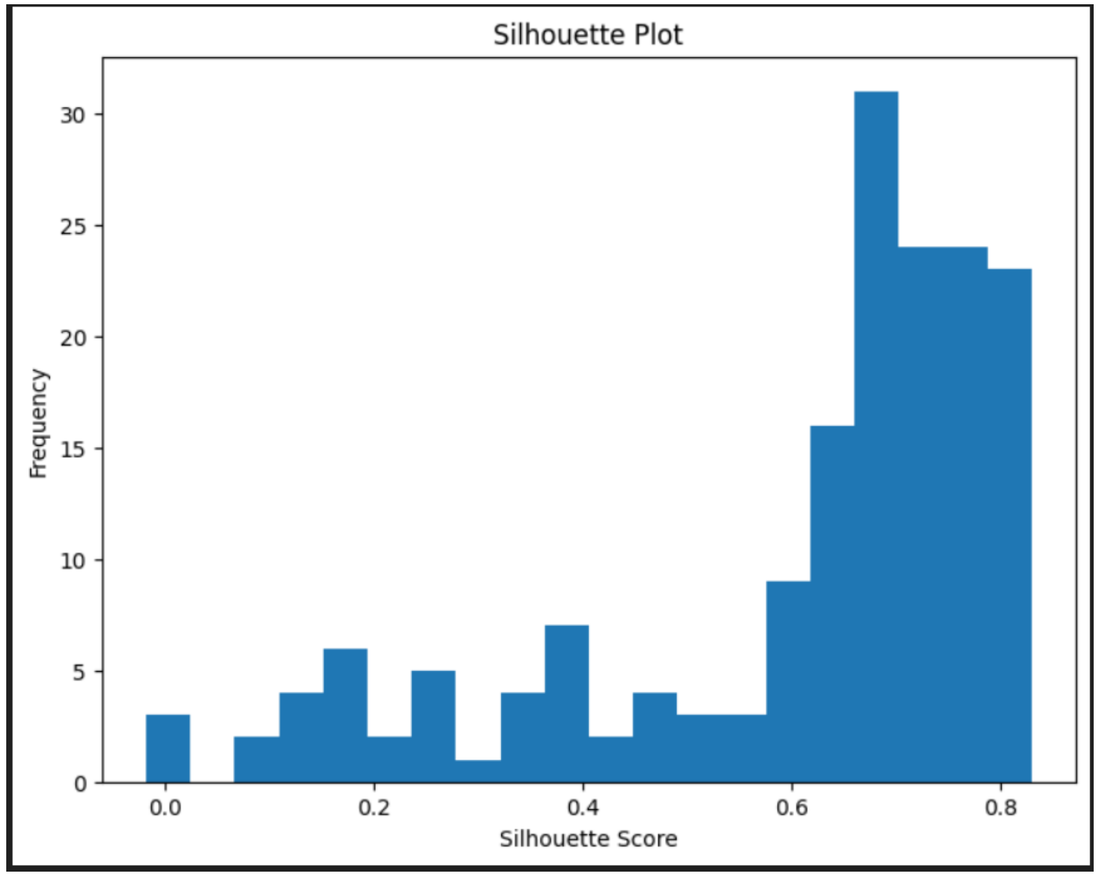
It ranges from -1 to 1, where a high value indicates good clustering

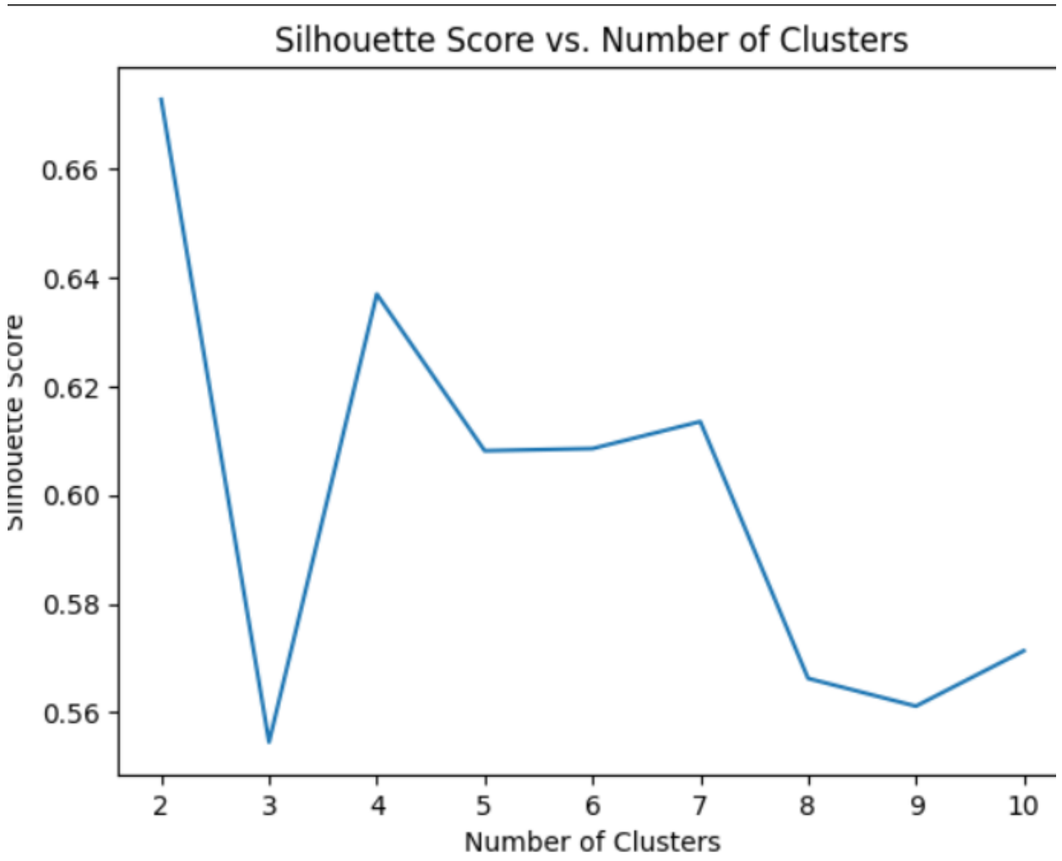
For our model:

Silhouette Score: 0.6081457349172995

Average Silhouette Score: 0.5817398863604938

Silhouette Plots





Silhouette Plots

These plots contains the silhouette scores corresponding to number of clusters

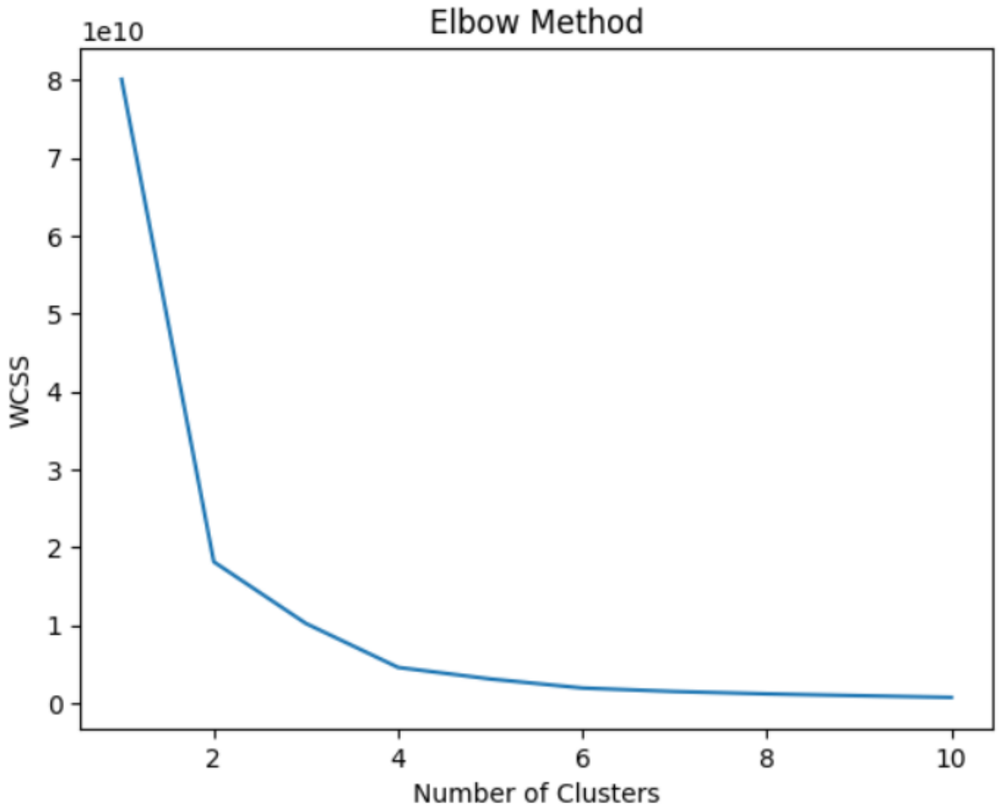
Higher silhouette value implies good clustering

Inferences:

The looks good enough as silhouette histogram contains high frequency of points in high silhouette value region

From the Silhouette vs number of clusters, it can be said that optimal number of clusters is 4

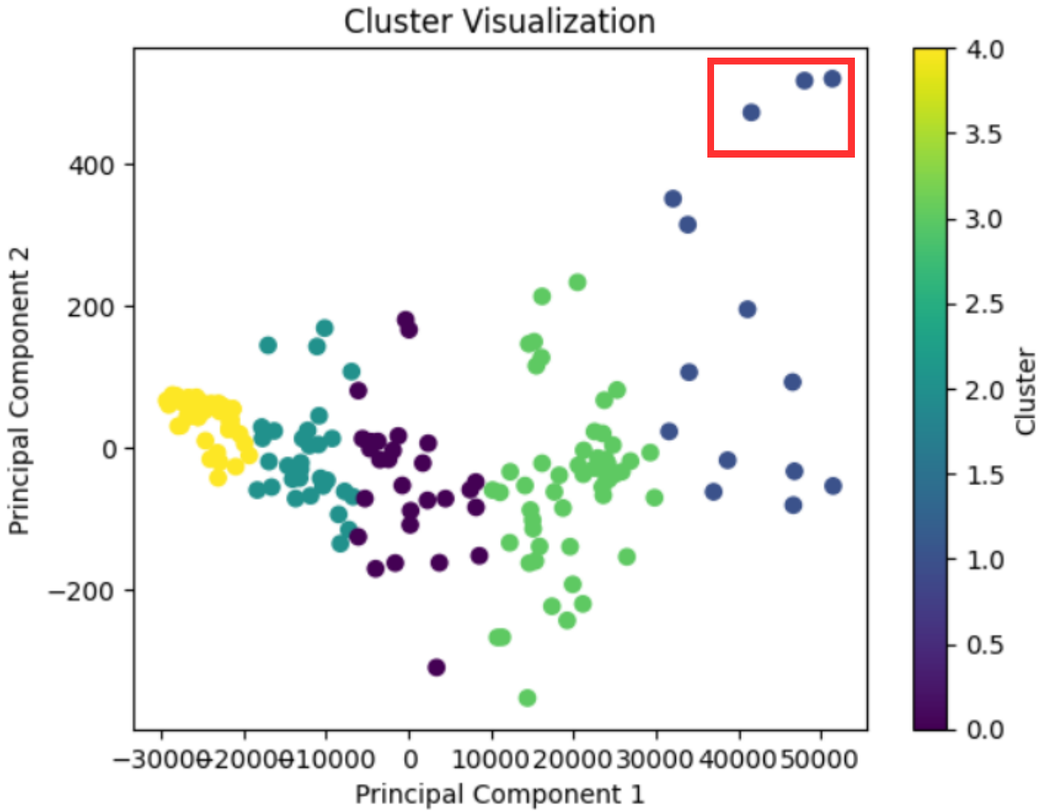
Elbow Plot



Elbow plots are another way to find optimal number of clusters

From the Elbow plots, it can be said that optimal number of clusters is 4

Cluster Visualization



helps in visualizing clusters in 2D space using PCA

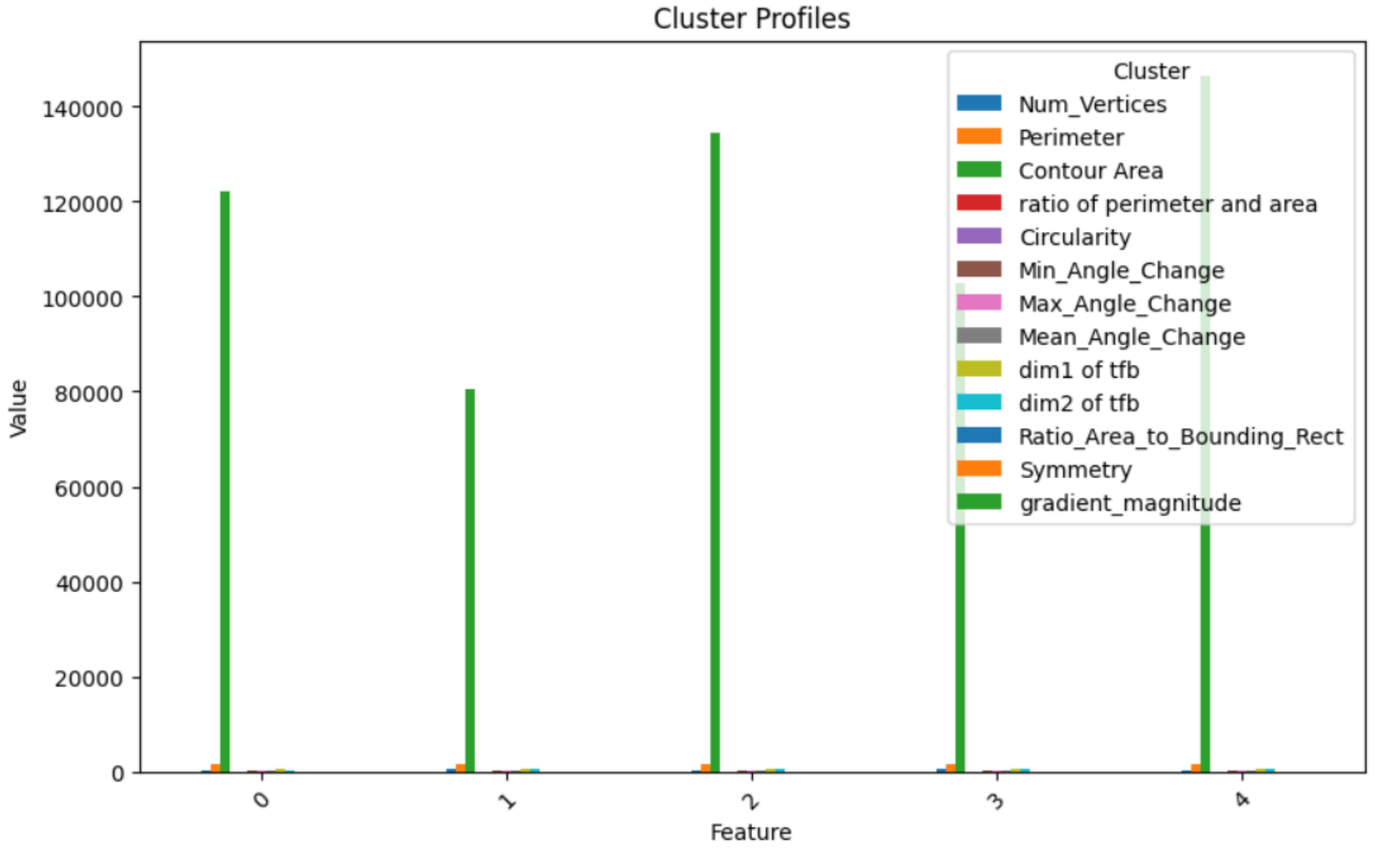
Inference:

Well-separated clusters, indicates features used for clustering are effective

Non-overlapping clusters

A few outliers can be seen, as highlighted

Feature Comparison

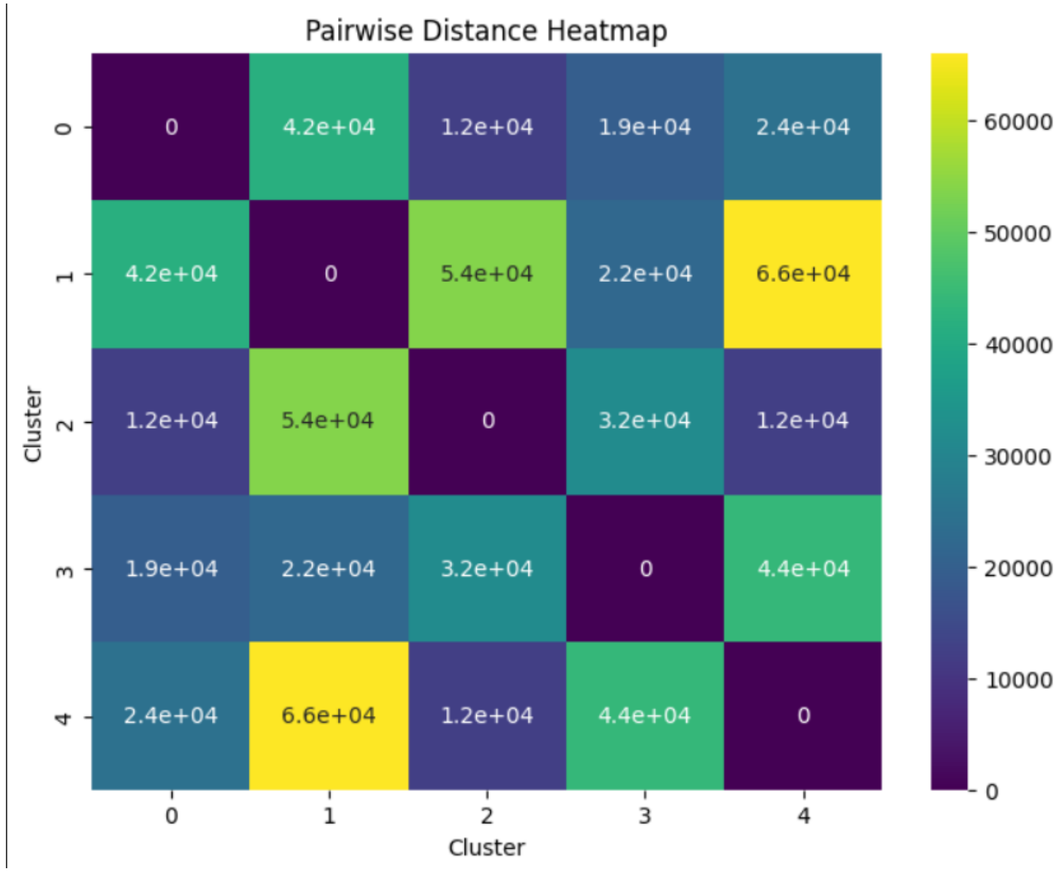


helps in comparing the contribution of each feature

Average Values:	
Num_Vertices	283.045711
Perimeter	1496.313470
Contour Area	117334.388857
ratio of perimeter and area	0.013318
Circularity	0.656918
Min_Angle_Change	76.433036
Max_Angle_Change	135.000000
Mean_Angle_Change	132.788189
dim1 of tfb	413.959510
dim2 of tfb	365.615674
Ratio_Area_to_Bounding_Rect	0.771755
Symmetry	7.342875
gradient_magnitude	0.695635

Gradient magnitude has the highest contribution

Pairwise Distance Heatmap



helps in analyzing similarity between clusters

Higher distance value means that clusters are less similar

Cluster 1 and Cluster 4 are most dissimilar followed by the pair 1,2 and 3,4

4 Clusters Formed After Analysis


0014
0023
0028
0034
0047
0051
0056
0057
0060
0062
0066
0072
0076
0077
0079
0081
0105
0111
0113
0118
0126
0129
0131
0148
0152
0203
0204
0209
0211
0226
0232
0233
0262
0315
0382
0403
0413
0424
0429
0445
0451
0459
0467
0521
0523
0551

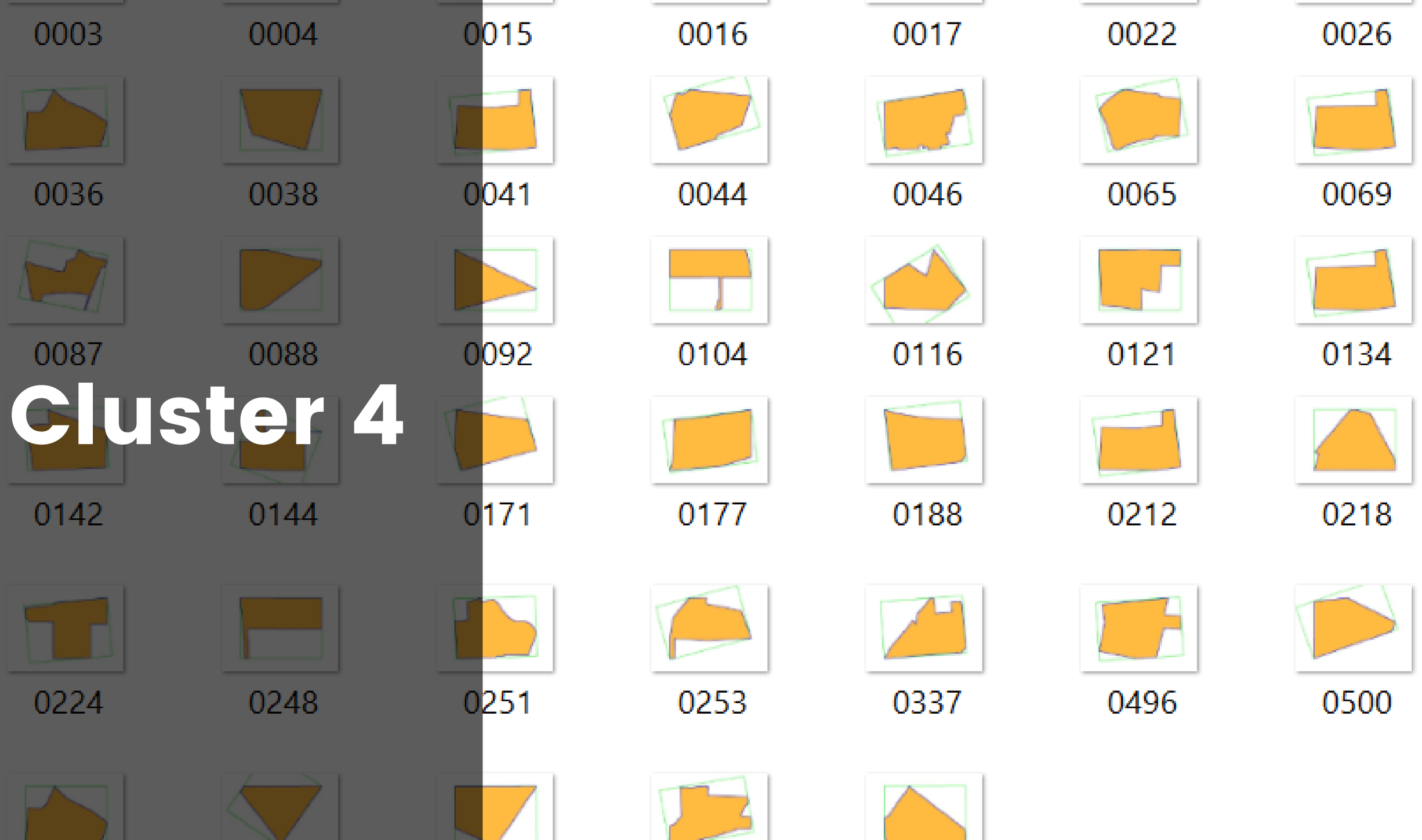
Cluster 1


0031
0040
0042
0045
0048
0052
0099
0112
0122
0165
0180
0197
0238
0239
0264
0318
0433
0503
0607
0619
0687
0971

Cluster 2


0011
0020
0029
0030
0032
0033
0035
0037
0039
0050
0059
0067
0074
0075
0095
0109
0114
0115
0119
0140
0146
Cluster 3
0166
0181
0198
0227
0236
0249
0252
0274
0275
0287
0310
0347
0352
0358

Cluster 4


0003
0004
0015
0016
0017
0022
0026
0036
0038
0041
0044
0046
0065
0069
0087
0088
0092
0104
0116
0121
0134
0142
0144
0171
0177
0188
0212
0218
0224
0248
0251
0253
0337
0496
0500

Design Family formation

Method 2 : CNN

Feature Extraction - VGG16

VGG16 is a pre-trained model to extract deep features from images.These features are extracted from 'block4_conv3' layer, known for capturing rich visual representations

K-means clustering

Using these features, we performed K-means clustering algorithm to form required number of groups of images

4 Clusters Formed By CNN Method

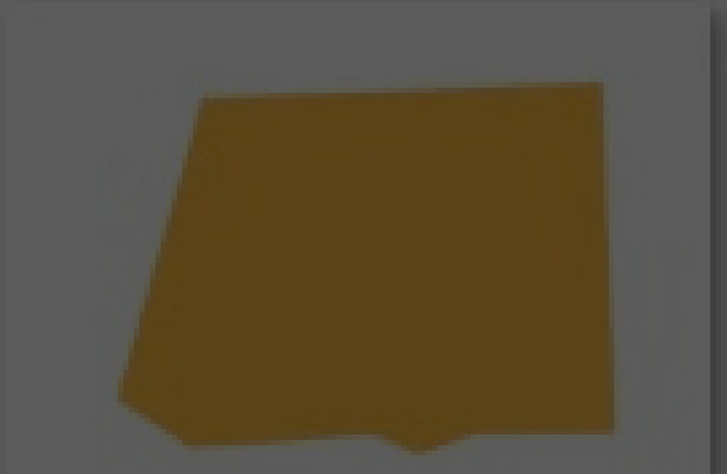

Cluster 1

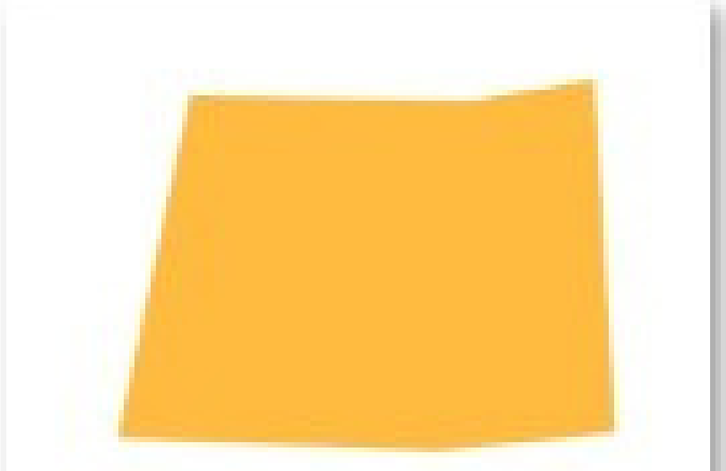
Cluster 2

0011

0012

0047






Cluster 3
0003
0004
0006
0009
0013
0016
0020
0021
0023
0024
0025
0026
0034
0039
0041
0043
0044
0045
0048
0049
0055
0059
0067
0068
0069
0073
0082
0092
0093
0096
0099
0100
0103
0106
0107
0110
0111
0113
0122
0124
0125
0131
0133
0143
0144
0147
0148
0149
0151
0157
0160

Cluster 4

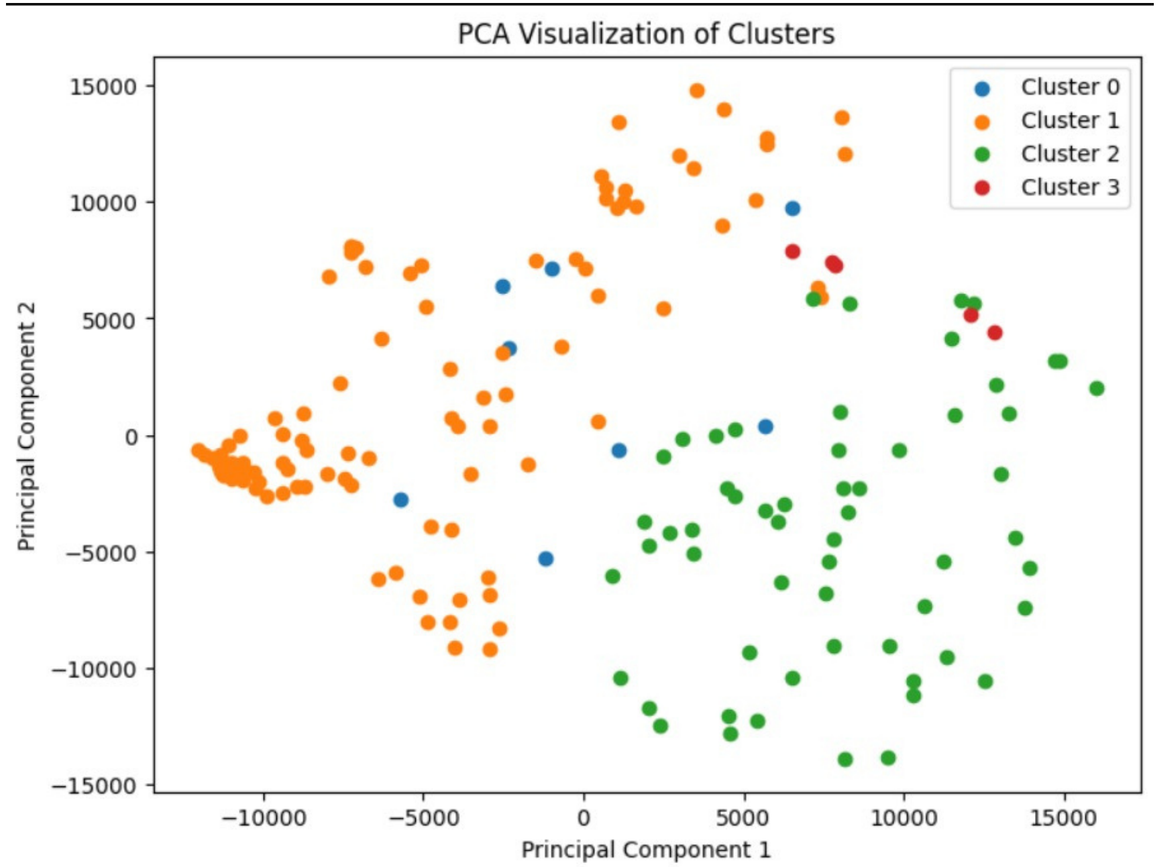
0008

0029

0072

Model Evaluation

Cluster Visualization



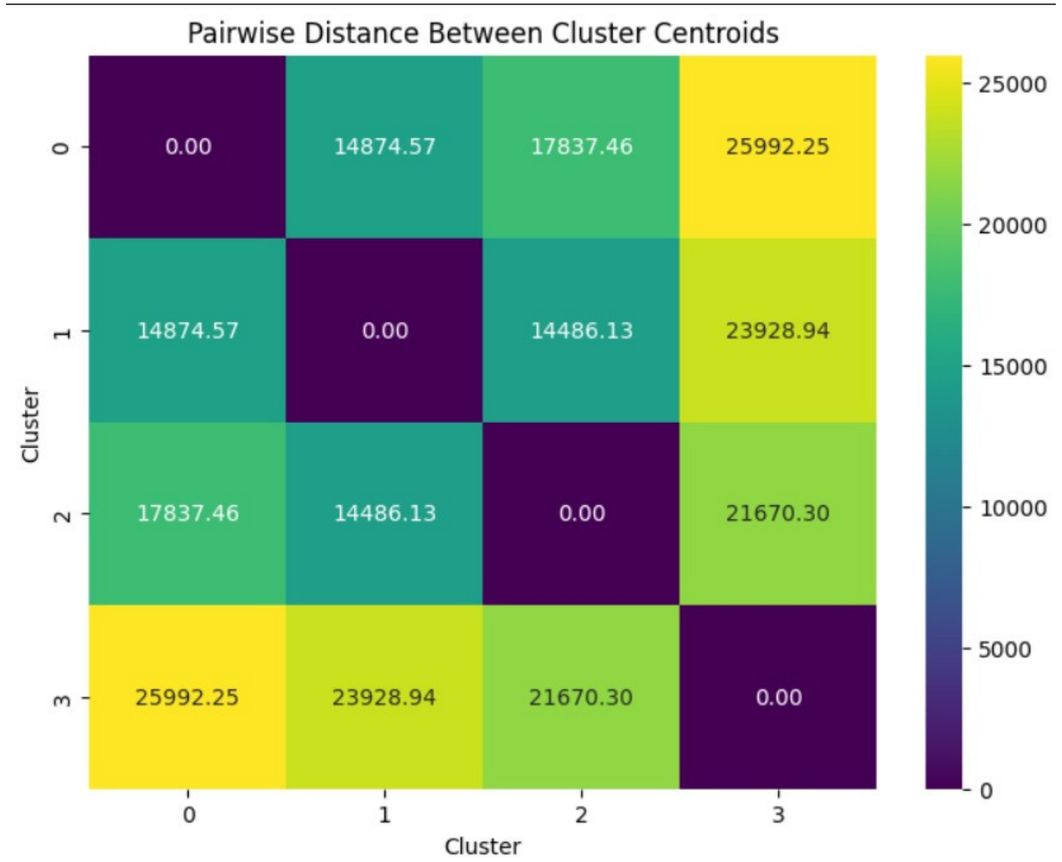
helps in visualizing clusters in 2D space using PCA

Inference:

Cluster 1 and 2 are more dense as compared to cluster 0 and 3

They have some overlapping among them

Pairwise Distance Heatmap



helps in analyzing similarity between clusters

Higher distance value means that clusters are less similar

Cluster 0 and Cluster 3 are most dissimilar followed by the pair 1,3 and 2,3

Model Comparision

Silhouette Score

Method 1

Silhouette Score:
0.6081457349172995

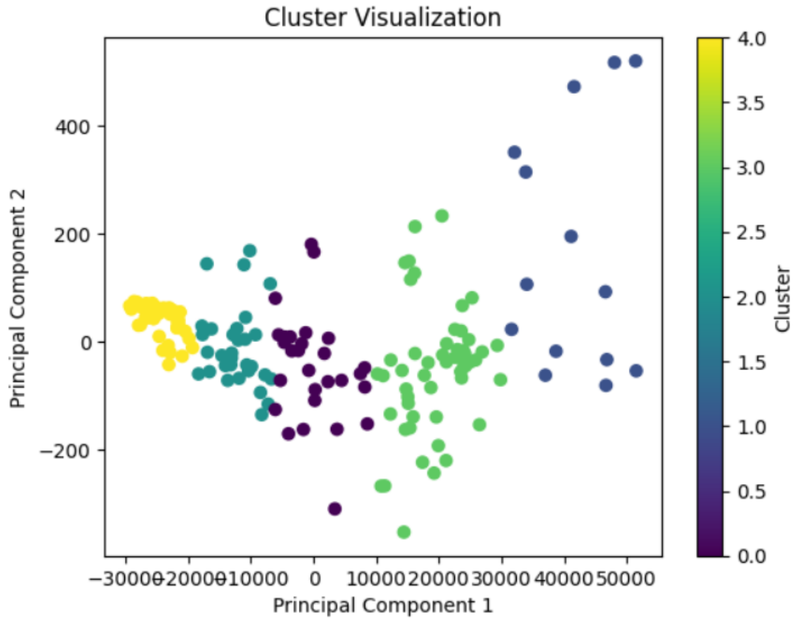
Method 2

Silhouette Score:
0.08565911

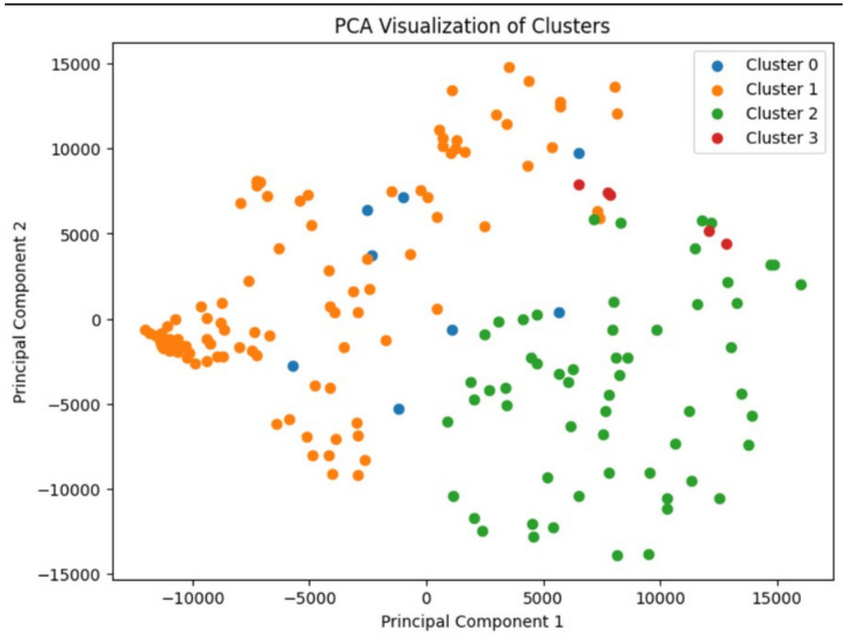
Method 1 have a significantly better silhouette score as compared to method 2, which is CNN., which indicates that method 1 is far better than method 2

Cluster Visualization

Method 1



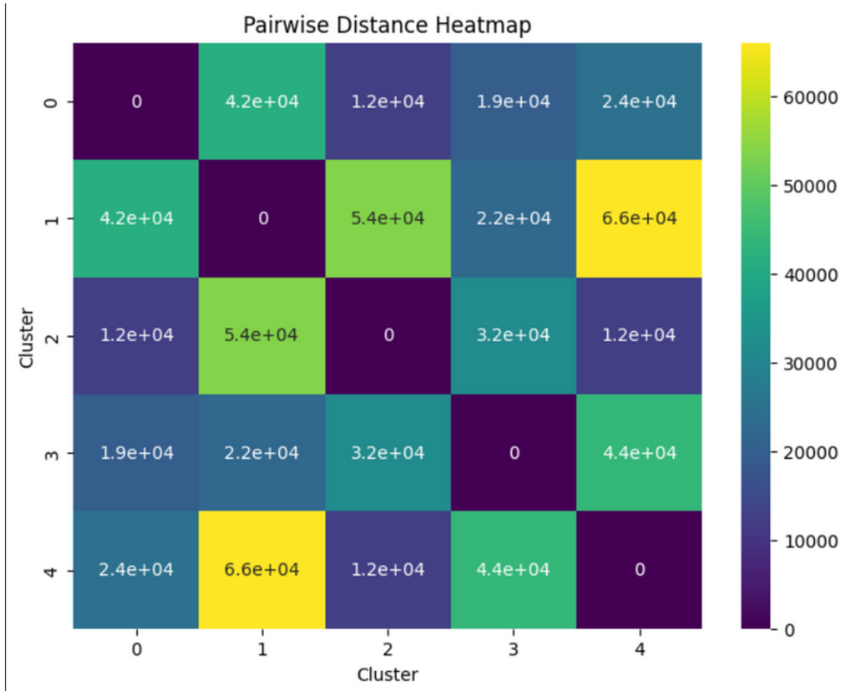
Method 2



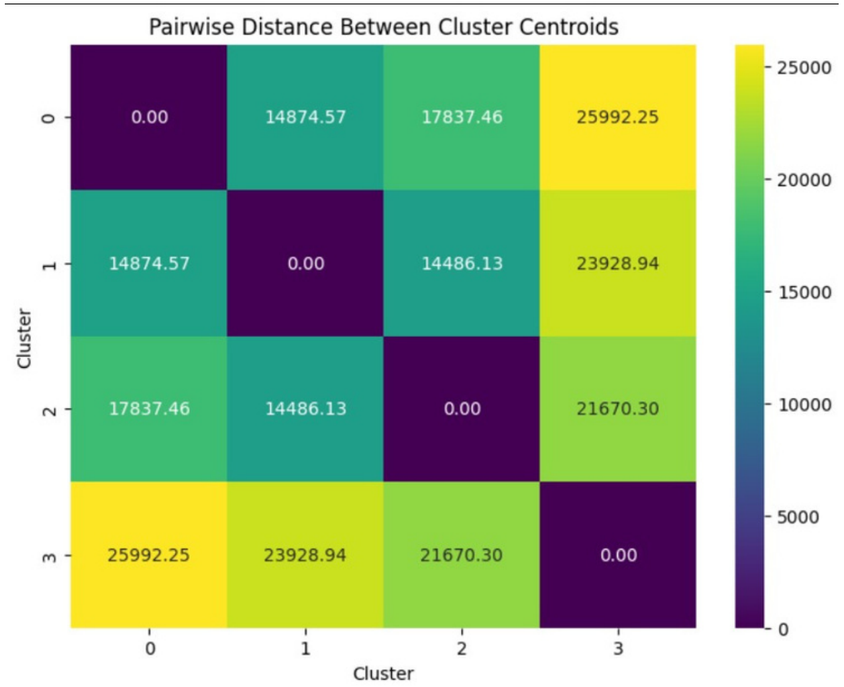
Method 1 have better, well-seperated, non-overlapping, and evenly distributed clusters as compared to method 2

Pairwise Distance Heatmap

Method 1



Method 2



Most of the clusters formed by method 1 have low distance value, which implies medium or high similarity, whereas most of the clusters formed by method 2 have high distance value, which implies low similarity

Complexity Classification

Features used:
Ratio of perimeter and area
Circularity
Min Angle Change
Max Angle Change
Mean Angle Change
Ratio Area to Bounding Rectangle
Symmetry
Gradient magnitude

We defined thresholds based on the complexity measure for low, medium and high complexity

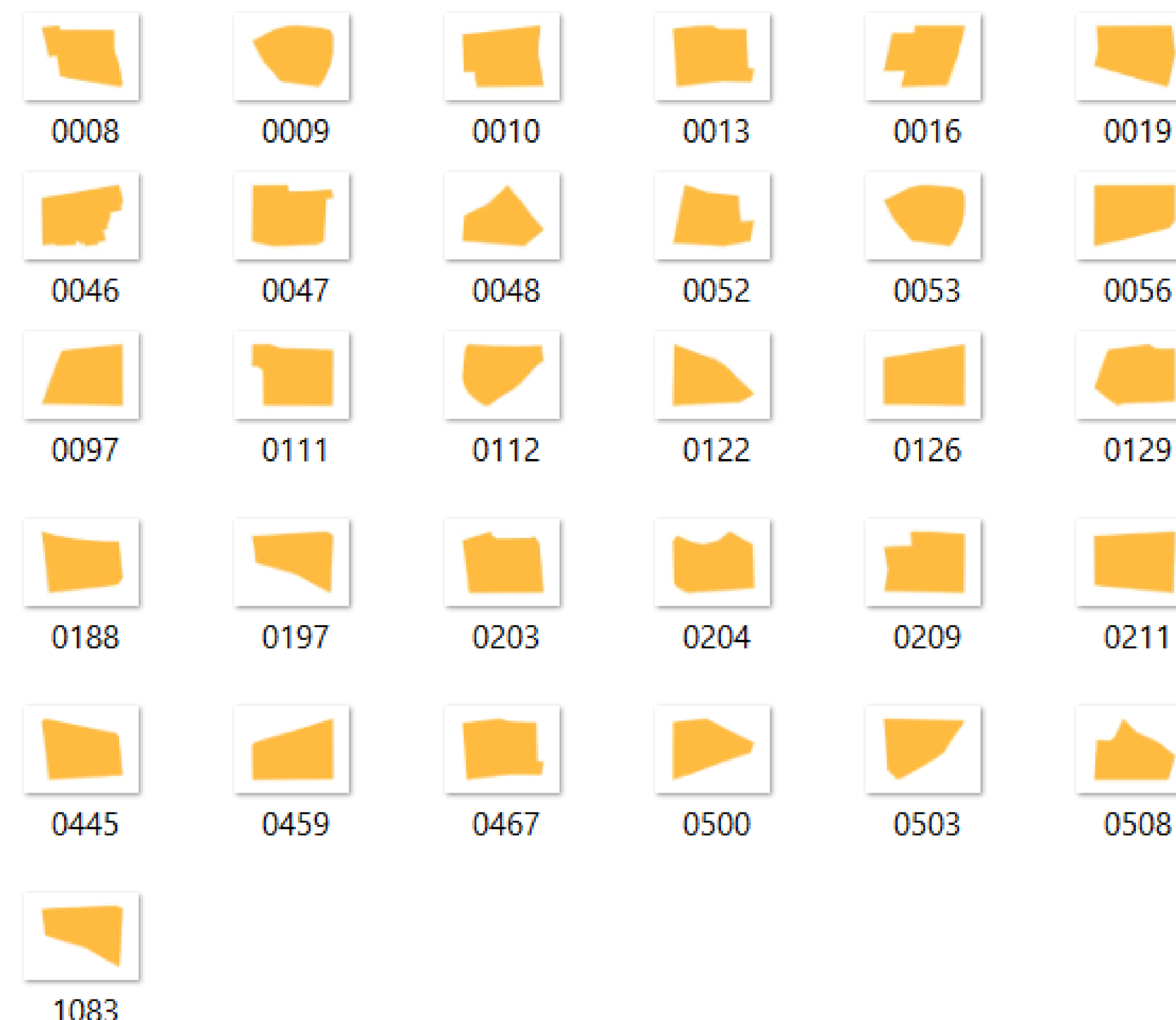
Then, similar to method 1, used for clustering, using the mentioned features, we performed PCA analysis to form optimum set of features for classification of images

Using these optimum set of features, we performed K-means clustering algorithm to form 3 groups, with low, medium and high complexities

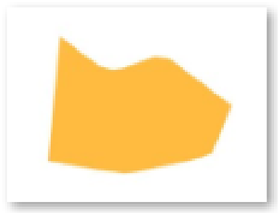
Comlexity Clusters

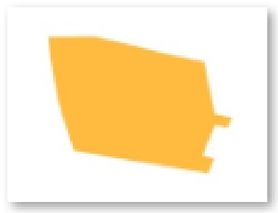
Low Complexity

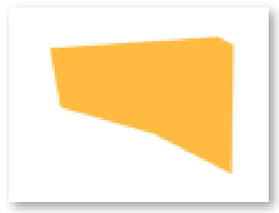
Medium Complexity

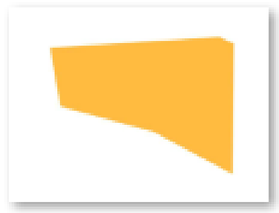


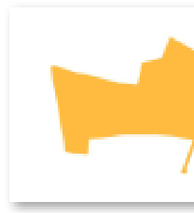
High Complexity

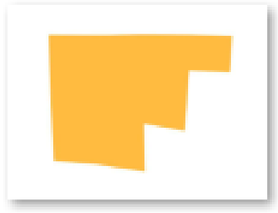


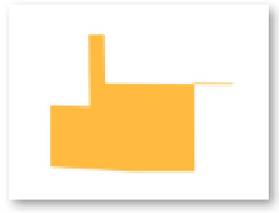


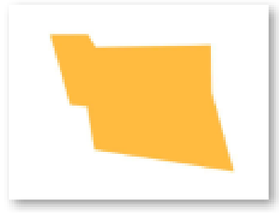


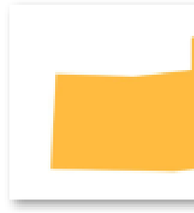


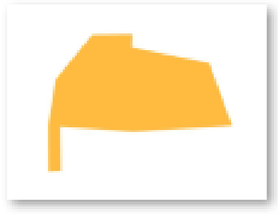


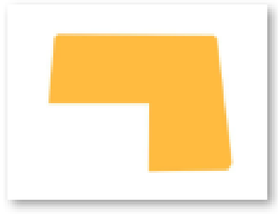


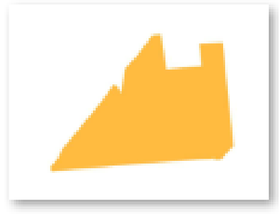


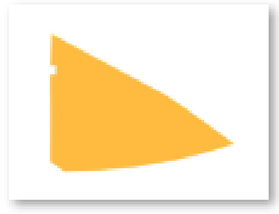


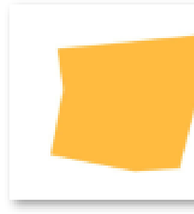












Layout Retrieval Model

Input from the user:

- dimensions (length and breadth) of tight-fitting box
- Complexity (low/ medium/high)
- Area of layout (or contour)

In this model, we are using the ratio of the dimensions for predicting relevant designs required, so as to eliminate the error we were facing earlier due to dimensions

Prediction of the Relevant output is performed using Nearest Neighbor method

Examples for Layout Retrieval Model

Example 1

Output

images of relevant layouts for dim1/dim2 ratio

images of relevant layouts for dim2/dim1 ratio

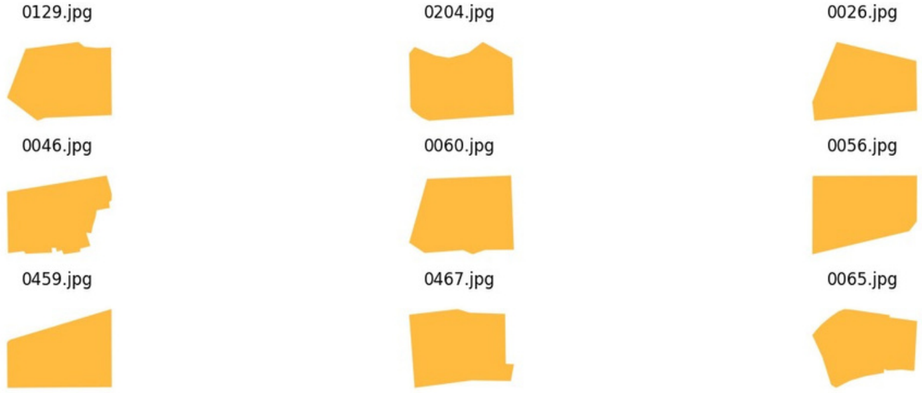
Input:

Dimension 1: 350
Dimension 2: 490
Complexity: 2 (high)
Area of layout: 120000

Example 2

Output

images of relevant layouts for dim1/dim2 ratio



images of relevant layouts for dim2/dim1 ratio

Input:

Dimension 1: 200
Dimension 2: 300
Complexity: 1 (medium)
Area of layout: 50000

Example 3

Output

images of relevant layouts for dim1/dim2 ratio

images of relevant layouts for dim2/dim1 ratio

Input:

Dimension 1: 350
Dimension 2: 470
Complexity: 0 (low)
Area of layout: 119000

Evaluation Criteria

Are all the questions posed by the problem statements effectively addressed with solutions?
Yes, we have tried to effectively find a solution for all the questions posed by the problem statement

Are the solutions relevant, and correctly applied, and backed up with proper logic / reasons?
We have tried to make our solution relevant and effective and tried to provide logic for required parts

Has there been any creative thinking and innovation while solving the problems?
Major part of creative thinking and innovation was involved in the feature selection part of the work. We tried to gather every possible feature which may impact the clustering, and then eliminate the similar features

Evaluation Criteria

Have any possibilities, other than those asked for, been implemented, or listed in the presentation?
We tried a few other possible methods for grouping, but they have nit been listed in the presentation

Are the major steps of data analysis diligently followed and correctly applied and documented (wherever required ...)?
We performed data analysis, along with manual visual inspection of data

Quality of results: are they backed with appropriate metrics, comparisons, analysis, explanations, and justifications?
We have performed analysis to check the quality of the model and results and also compared the results of two models when required

Evaluation Criteria

Quality of presentation: Completeness and preciseness of the final slide deck; design and readability of the slides. Are all the above questions covered in the presentation?

We tried to keep the presentation concise, readable and appealing and also tried to cover all the questions that were to be covered in the presentation

Does the presentation contain raw and hyped-up outputs generated using LLM tools like ChatGPT / Gemini etc.

We have not included any raw output generated using tools like ChatGPT etc. We have just used ChatGPT to help us with the coding part and to understand certain things

Problems That We Faced While Doing The Project

We faced problems to find methods that should be used to form clusters based on the shapes

Feature engineering was a big challenge for us Choosing right and unique features which would aid in better clustering, required thinking from an archtitect point of view

We were not sure about the optimum number of features that should be obtained using PCA

During building the model for retrieving layouts based on certain parameters, the image/layout was 640:480 pixels ratio, while the input being provided can be of any dimension or order.
We needed to find a way to incorporate any order value and give required output

Thank You :)