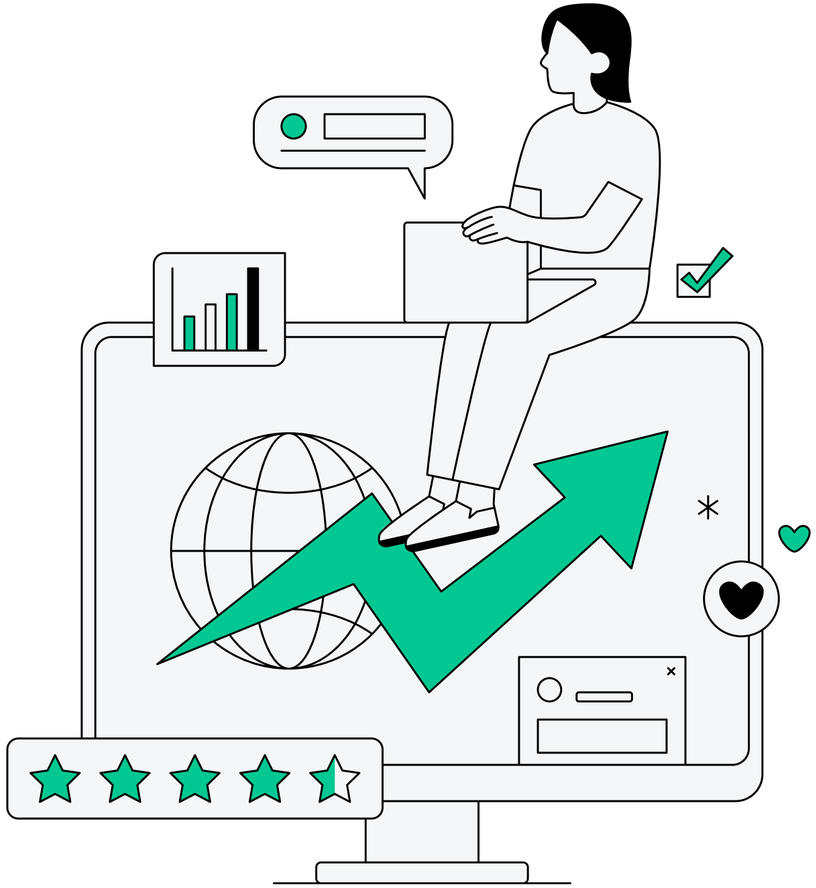
Year/Semester: 2023-24/Spring

Course Code	Course Name	Credits	Tag	Grade	Credit/Audit
CL 208 (S2)	Chemical Reaction Engineering	6.0	Core course	BC	C
CL 210	Separation Processes	6.0	Core course	CC	C
CL 232	Chemical Engineering Lab. I	6.0	Core course	AB	C
CL 238 (S2)	Introduction to Numerical Analysis	6.0	Core course	BC	C
CL 242	Fundamentals of Heat and Mass Transfer	6.0	Core course	BB	C
DE 250 (S1)	Design Thinking for Innovation	6.0	Core course	AA	C
DS 203 (M)	Programming for Data Science	6.0	Minor	BC	C
NOCS01	NCC/NSS/NSO	0.0	Core course	Not allotted	N

DS 203
E7 Project

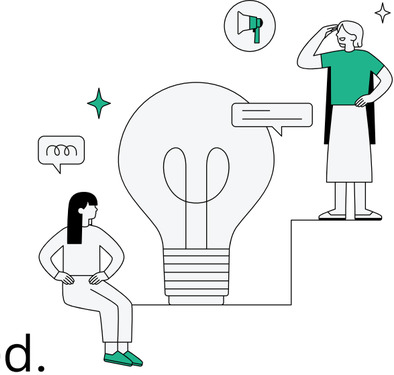
- Afnan Abdul Gafoor (22B2505)
- Shaheem Basheer (22B0446)



Problem Description

- An architecture company has digitized their building layout designs into 640 x 480 pixel bitmaps, creating a repository of 1183 such views. They aim to leverage data science and machine learning to improve productivity, insights, and responsiveness to customer needs.
- Categorize designs into families based on shapes for insights and template creation. Explore multiple approaches for effective classification.
- Classify layouts into Low, Medium, or High complexity through formal analysis.
- Retrieve similar layouts based on specified parameters like dimensions and complexity. Predict design families for efficient layout retrieval, saving time and ensuring quality.
- Use image data for innovative enhancements in design processes and productivity.

Objective



Grouping Designs into Families:

- Explore methods to group designs based on their shapes into distinct families.
- Multiple approaches will be tested, with results from at least two approaches presented.

Classifying Layout Complexity:

- Analyze layouts to classify complexity into Low, Medium, or High categories.
- Establish criteria for complexity classification based on formal analysis of layouts.

Speeding up Layout Design Process:

- Develop a system to retrieve relevant prior layouts based on given parameters.
- Parameters include dimensions (length, width) of the tight-fitting box, layout area, and permissible complexity.
- Predict design family/families closest to specified parameters for efficient retrieval.

Exploring Further Possibilities:

- Solicit suggestions for additional tasks leveraging image data and mined information.
- ideas on innovative applications or analyses beyond the specified tasks.

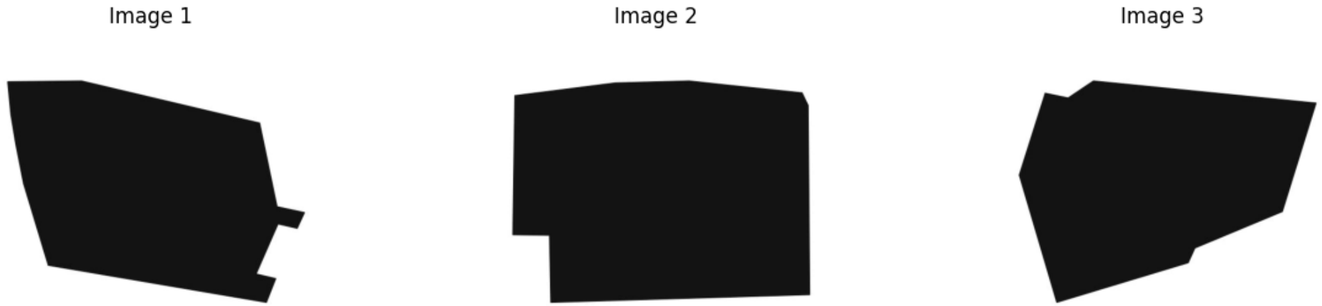
Summary

- The dataset is loaded, and preliminary exploratory data analysis (EDA) is conducted to gain insights into its characteristics.
- Our exploration involved generating numerous shape-related features using OpenCV, a powerful library for image processing. Through this process, we gained insights into the various characteristics and properties of shapes, enabling us to analyze and understand images in greater detail..
- Engineered features are stored in a CSV file for organized storage and future reference.
- A correlation heatmap is generated to visualize relationships between different features, aiding in identifying significant correlations.
- Dimensionality reduction techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE), are employed to reduce the dataset's dimensionality while preserving essential relationships.
- Two clustering algorithms, K-Means and hierarchical clustering, are applied to group layouts into distinct families based on shared characteristics.
- Layout complexity is quantified using the perimeter^2/Area ratio (solidity), providing a measure of intricacy and solidity in layout design.
- A user-friendly interface allows users to input parameters and retrieve similar layouts based on historical data, facilitating informed decision-making.
- Feature weights are assigned to various layout attributes in terms of dollars, enabling the prediction of estimated costs associated with each layout.

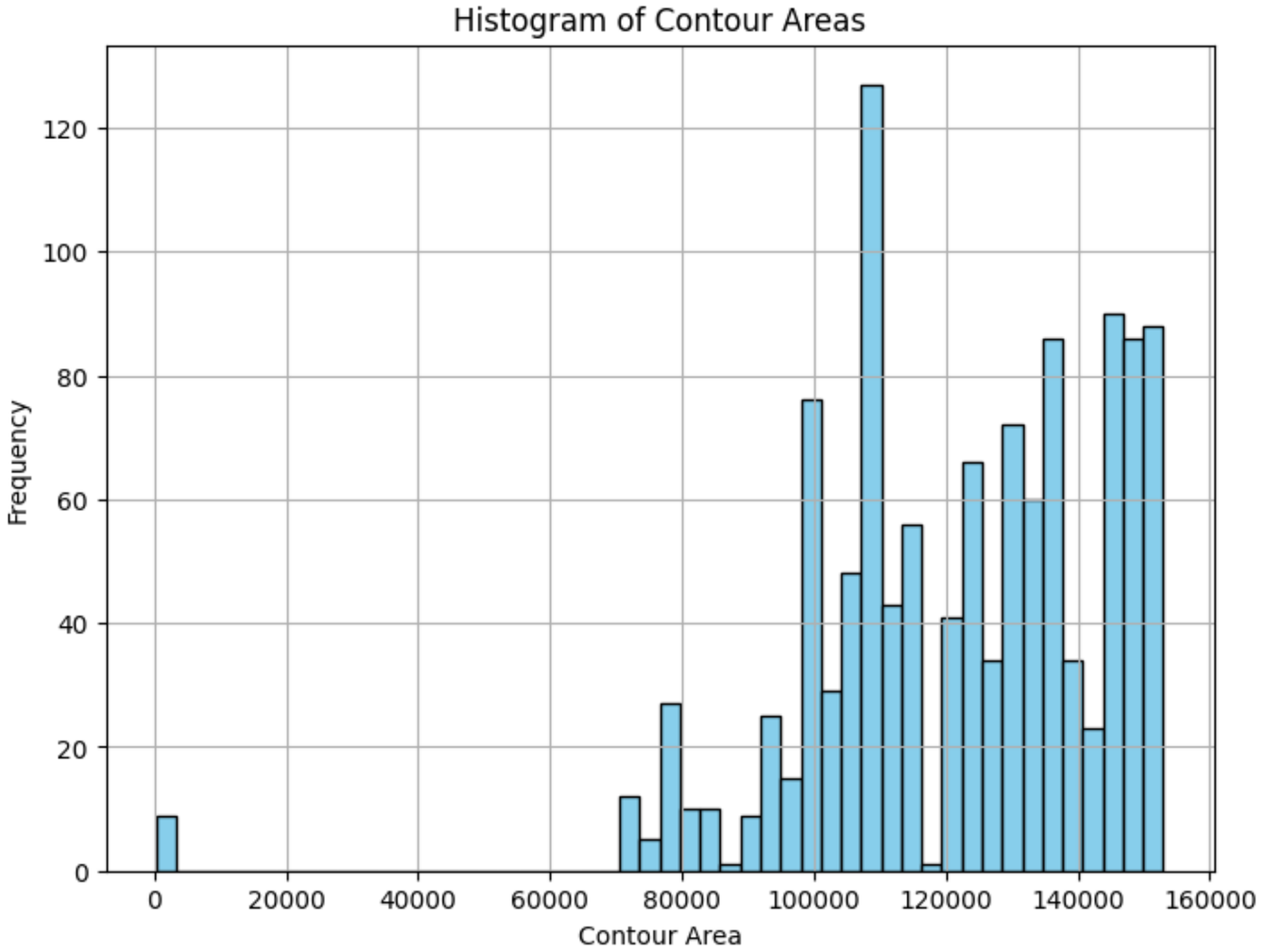
Data Observation

- The dataset consists of 1183 design layouts with dimensions and channels: (480, 640, 3)

- Sample greyscaled images:



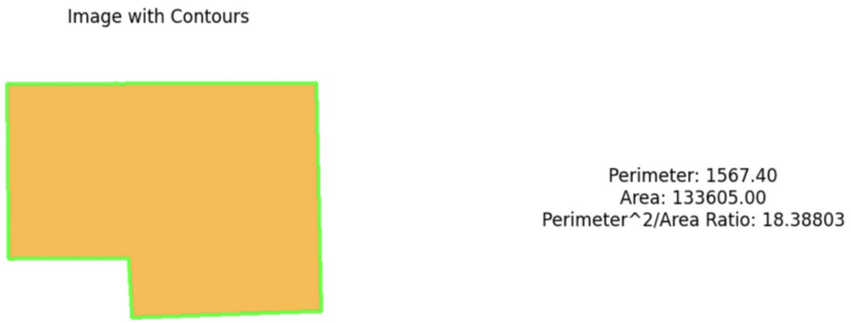
- The following area distribution for layouts is observed:



Creating features

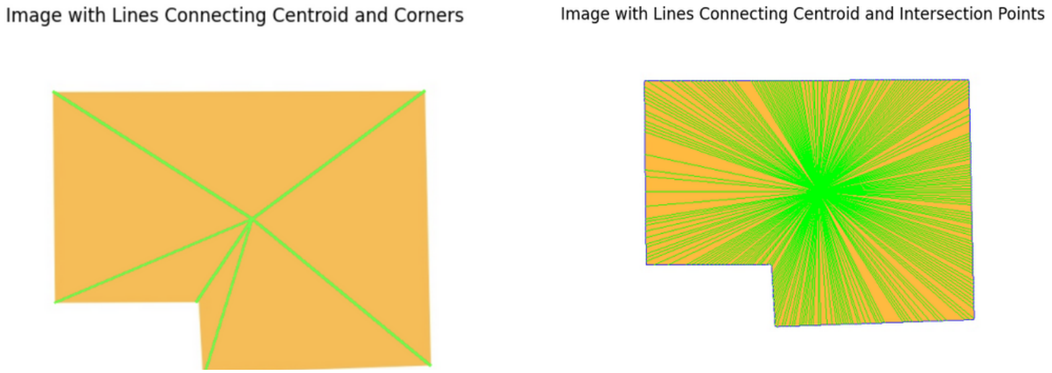
- Perimeter^2/Area Ratio

The ratio of the square of the perimeter to the area of the contour. It gives an indication of the shape complexity.



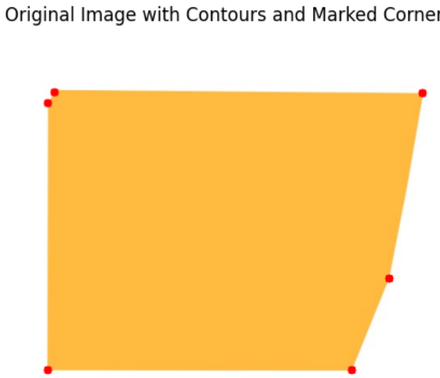
- Sum of distances from centroid

The sum of distances from the centroid of the contour to its intersection points with lines drawn from the centroid at various angles.



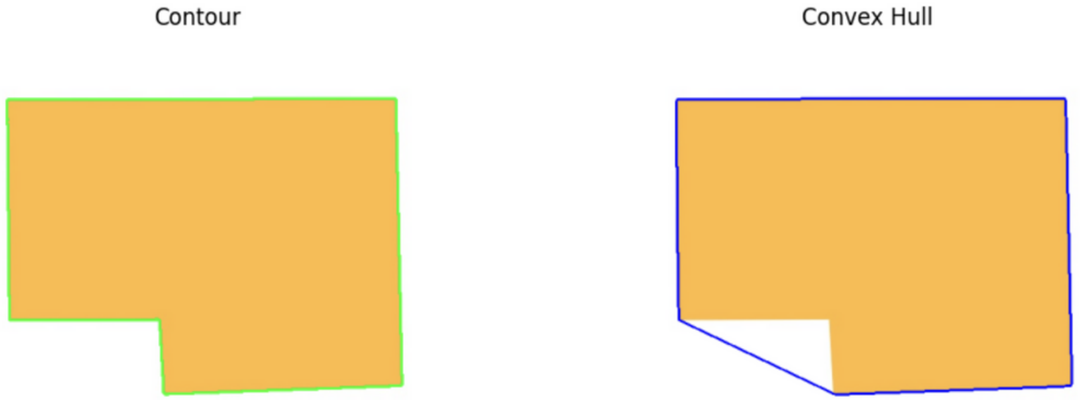
- # Total corners

The total number of corners detected in the contour.



- # Hull to contour area ratio

The ratio of the area of the convex hull to the area of the contour. It indicates how much the contour deviates from being convex.

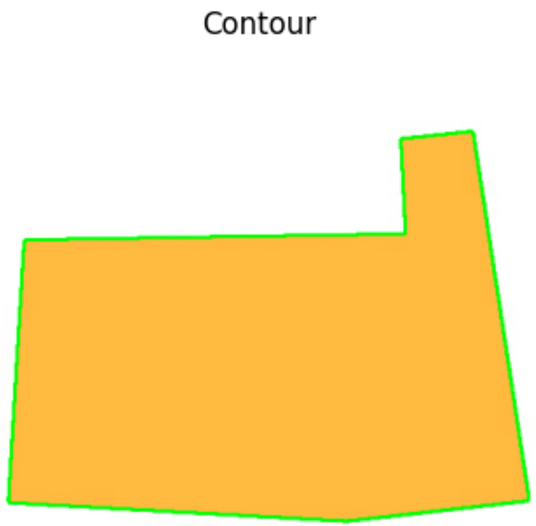


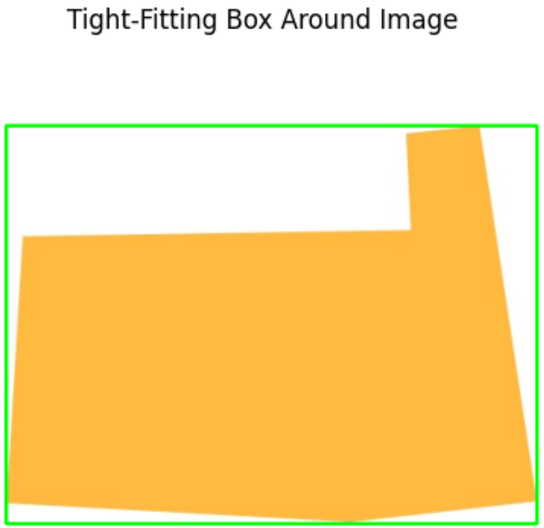
- Layout area

  Area of the contour, which represents the overall space occupied by the object in the image.

- Box to contour area ratio

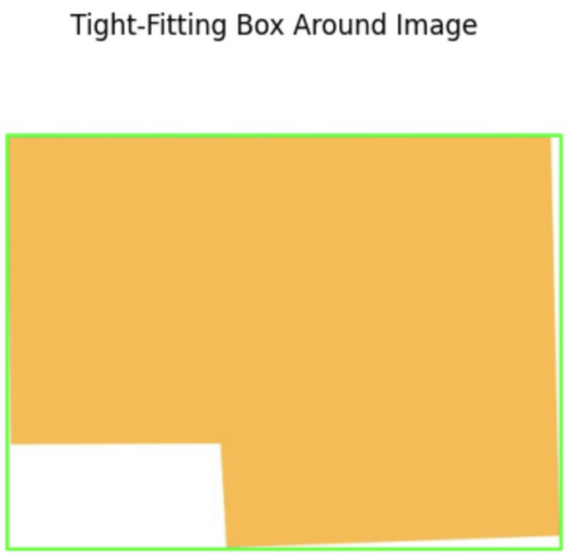
  The ratio of the area of the tight-fitting box (bounding rectangle) to the area of the contour.





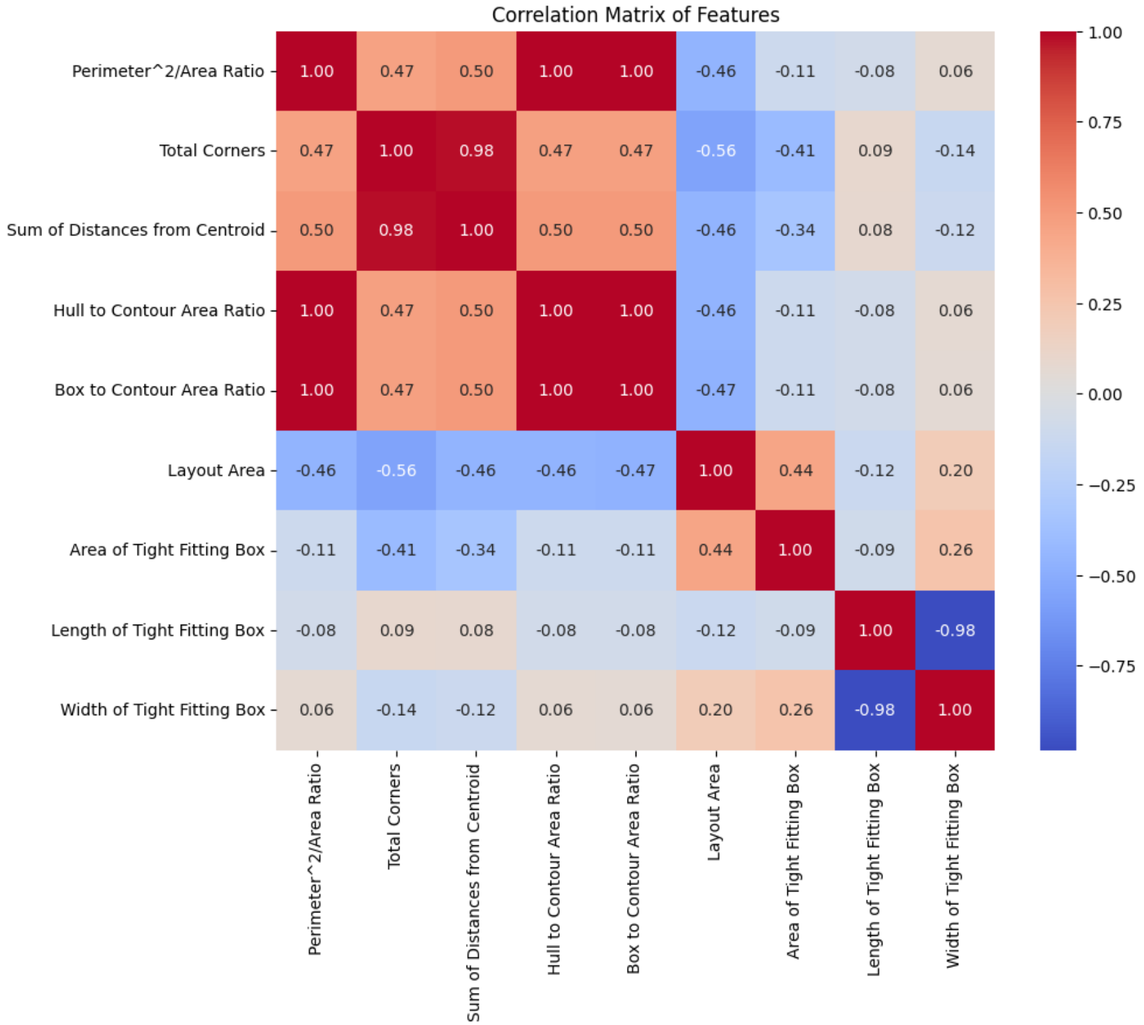
- Length & Width of Tight Fitting Box
- Area of Tight Fitting Box

These measurements collectively describe the size and shape of the tight-fitting bounding rectangle, providing insights into the spatial extent of the object represented by the contour.

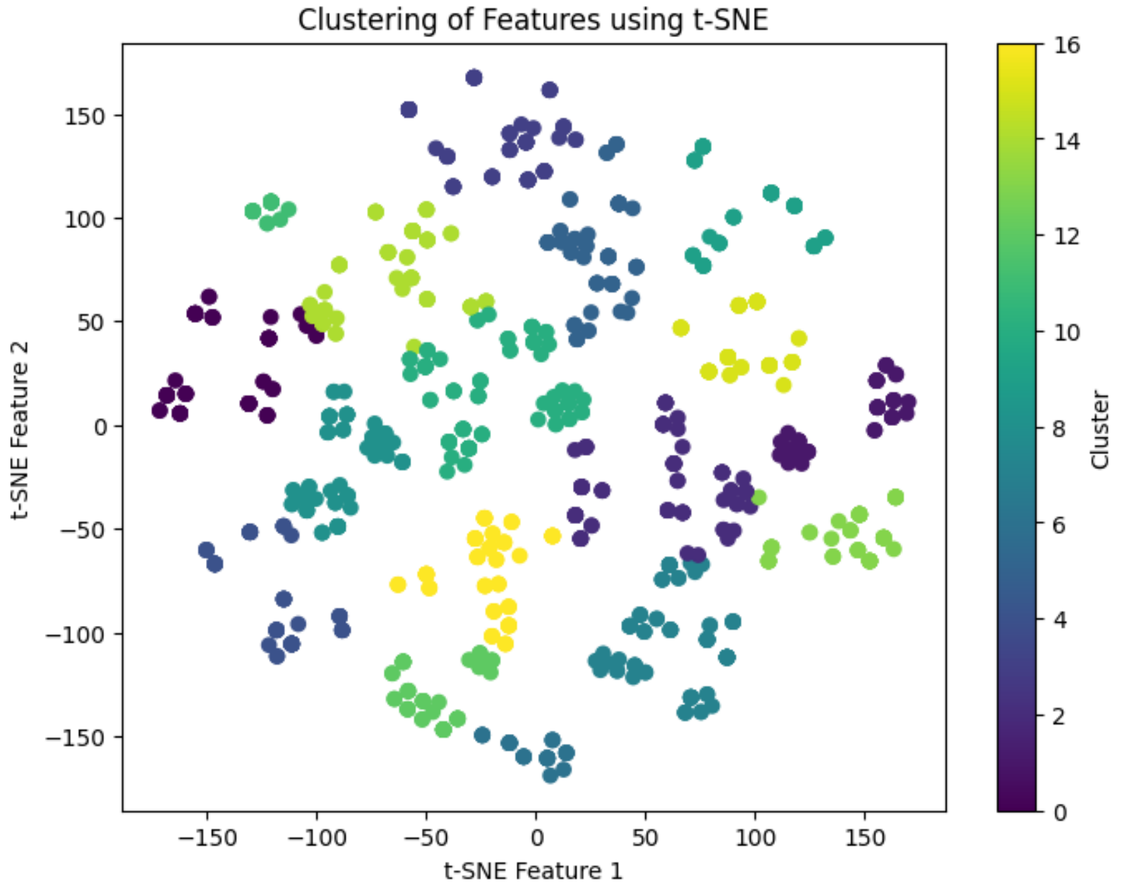


Family classification

- Using all available features in classification can lead to better results, even if some features appear correlated.

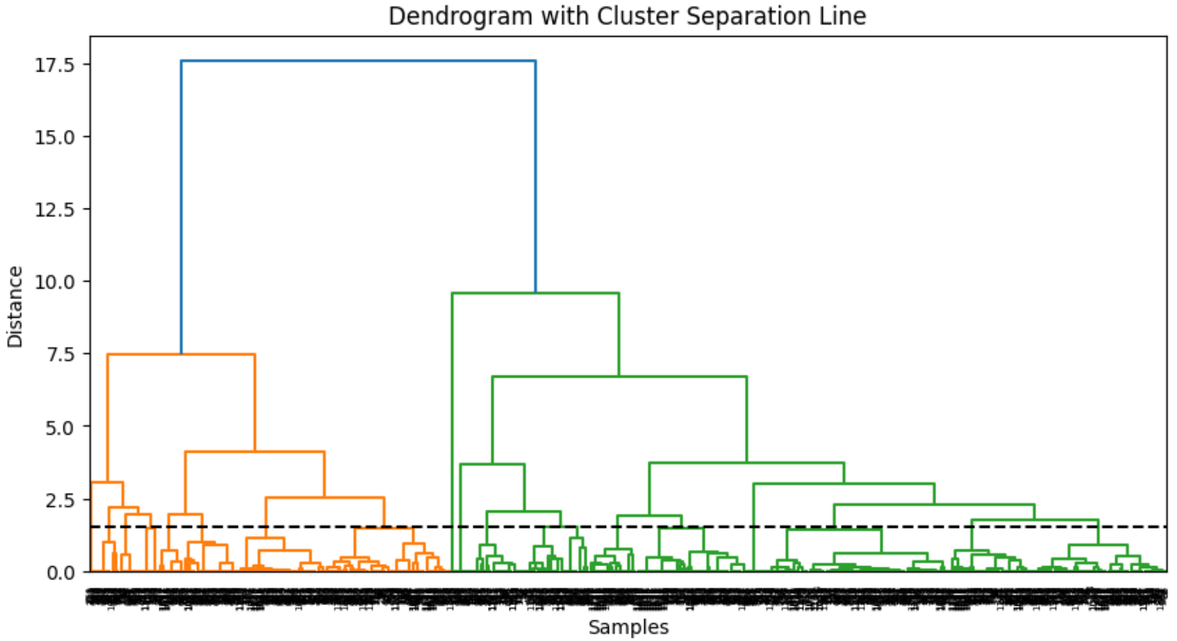


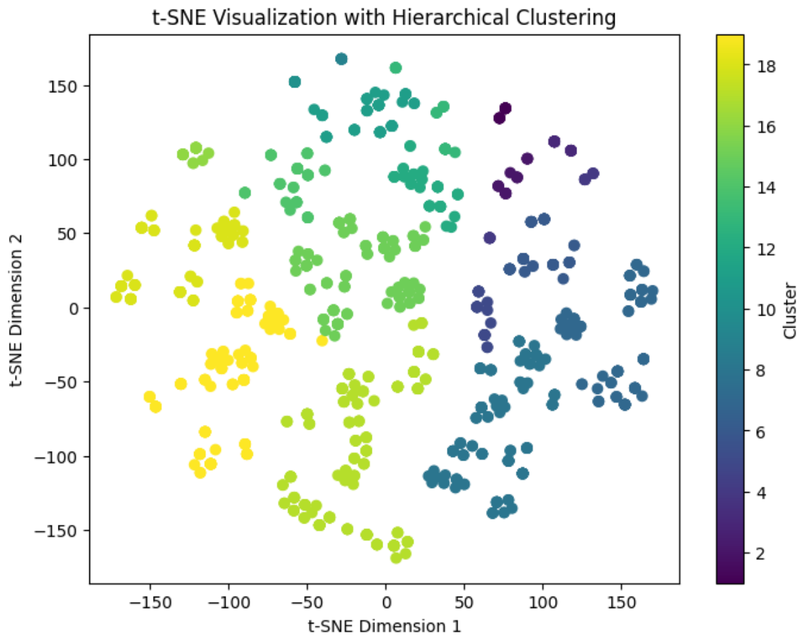
- We've loaded features from a CSV file, handpicked specific columns, scaled the features for consistency, utilized t-SNE for effective visualization, and executed K-means clustering to derive meaningful insights.



- The following clustering is obtained using K-means clustering
- 17 clusters
- This reveals similar designs across the rows.

- Exploring the dataset through dimensionality reduction using t-SNE, we visualize the data in a lower-dimensional space. We then apply hierarchical clustering to this reduced data and visualize it using a dendrogram. Setting a threshold of 1.5, we determine the number of clusters. Finally, we display the t-SNE visualization with cluster labels.





- The following clustering is obtained using dendrogram
- 19 clusters

Complexity classification

- Classifying layouts based on the perimeter^2/area ratio

Image: 0897.jpg
Complexity: Low Complexity

Image: 0224.jpg
Complexity: Medium Complexity

Image: 0526.jpg
Complexity: High Complexity

Predicting Design Families through Historical Data Analysis

- This code allows the architect to input desired parameters for a layout, such as the dimensions of the tight-fitting box, the layout area, and the complexity level. Based on these parameters, the code filters through a dataset of historical layout data and retrieves layouts that match the specified criteria. Finally, it displays the filtered layouts for the architect to review.

```
Enter the length of the tight-fitting box: 338
Enter the width of the tight-fitting box: 450
Enter the layout area: 138000
Enter the permissible layout complexity (0 for Low, 1 for Medium, 2 for High): 0
Enter the threshold for length: 10
Enter the threshold for width: 10
Enter the threshold for layout area: 10000
```

- Here are some images that match the specified criteria:

- These images showcase layouts with similar characteristics, providing a starting point for further exploration and refinement.

Exploring Further Possibilities with Image Data Analysis

- The script utilizes a set of predefined weights to calculate the cost of each layout based on parameters such as the length, width, and area of the tight-fitting box, as well as the complexity level. It then randomly selects layouts from a dataset and displays them alongside their estimated costs. This visualization provides users with insights into the cost implications of different layout designs, aiding in decision-making during the design process.

Our learnings & Possibilities

- **Feature Extraction from Images**: We gained insights into the process of extracting meaningful features from images, enabling us to leverage image data effectively in our analysis.
- **Dimensionality Reduction with t-SNE**: Exploring t-SNE, we learned how to effectively reduce the dimensionality of high-dimensional data while preserving its underlying structure, facilitating visualization and interpretation.
- **Importance of Domain Knowledge**: Recognizing the significance of domain knowledge, we understood how it enriches our analysis by providing context and guiding feature selection and interpretation.
- **Classification with K-Means and Hierarchical Algorithms**: Through classification tasks using K-Means and hierarchical algorithms, we acquired practical knowledge of clustering techniques and their applications in grouping data points based on similarity.