**Page Blocks Classification Dataset Analysis**

Usman Javed  
2022CS620

Data Science Course Project  
6th Semester

# Introduction

This report presents an analysis of the Page Blocks Classification dataset from the UCI Machine Learning Repository. The dataset consists of features extracted from document page layouts, with the goal of classifying different types of blocks (text, horizontal line, picture, vertical line, and graphic) in the page layout.

# Dataset Description

The dataset contains 5473 examples with 10 features and 1 target variable. The features include physical attributes of the blocks such as height, length, area, and various pixel-based measurements. The target variable represents five different classes of page layout elements.

# Univariate Analysis

The univariate analysis examines individual features independently to understand their distributions and characteristics.

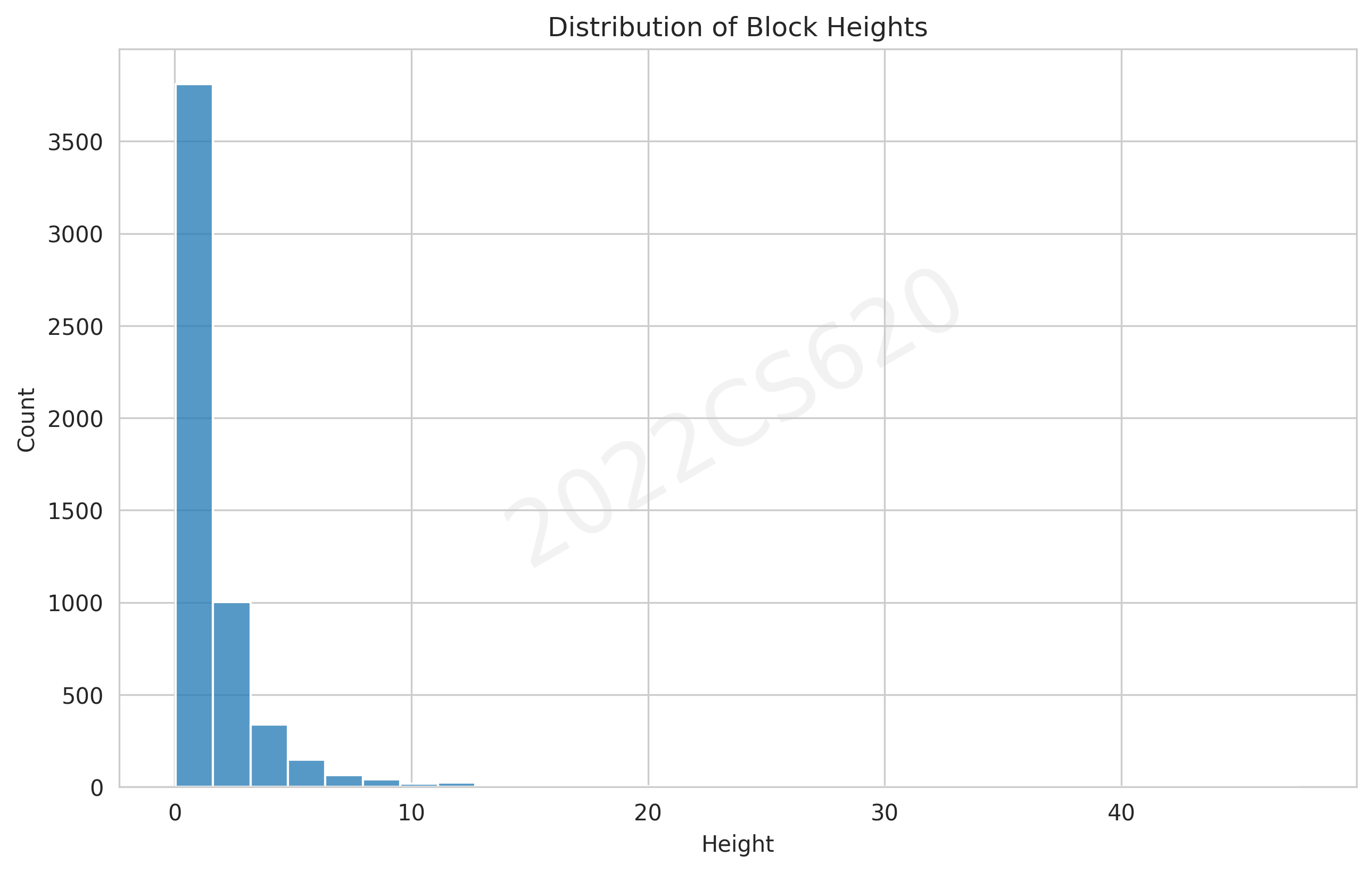


Figure 1: Distribution of block heights showing a right-skewed distribution with most blocks having small heights.

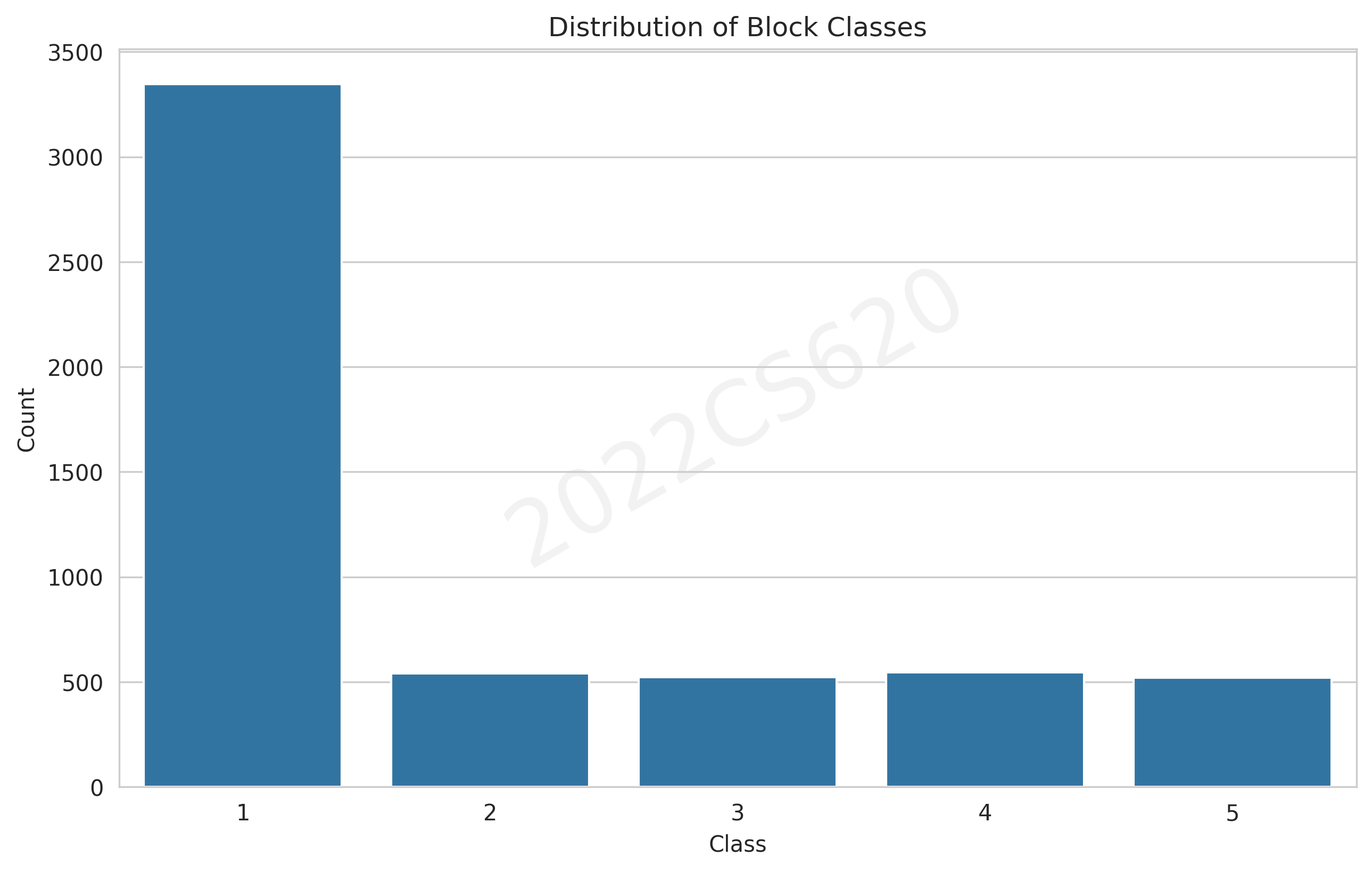


Figure 2: Distribution of block classes showing class imbalance with text blocks being the most common.

# Bivariate Analysis

The bivariate analysis explores relationships between pairs of features and their associations with the target class.

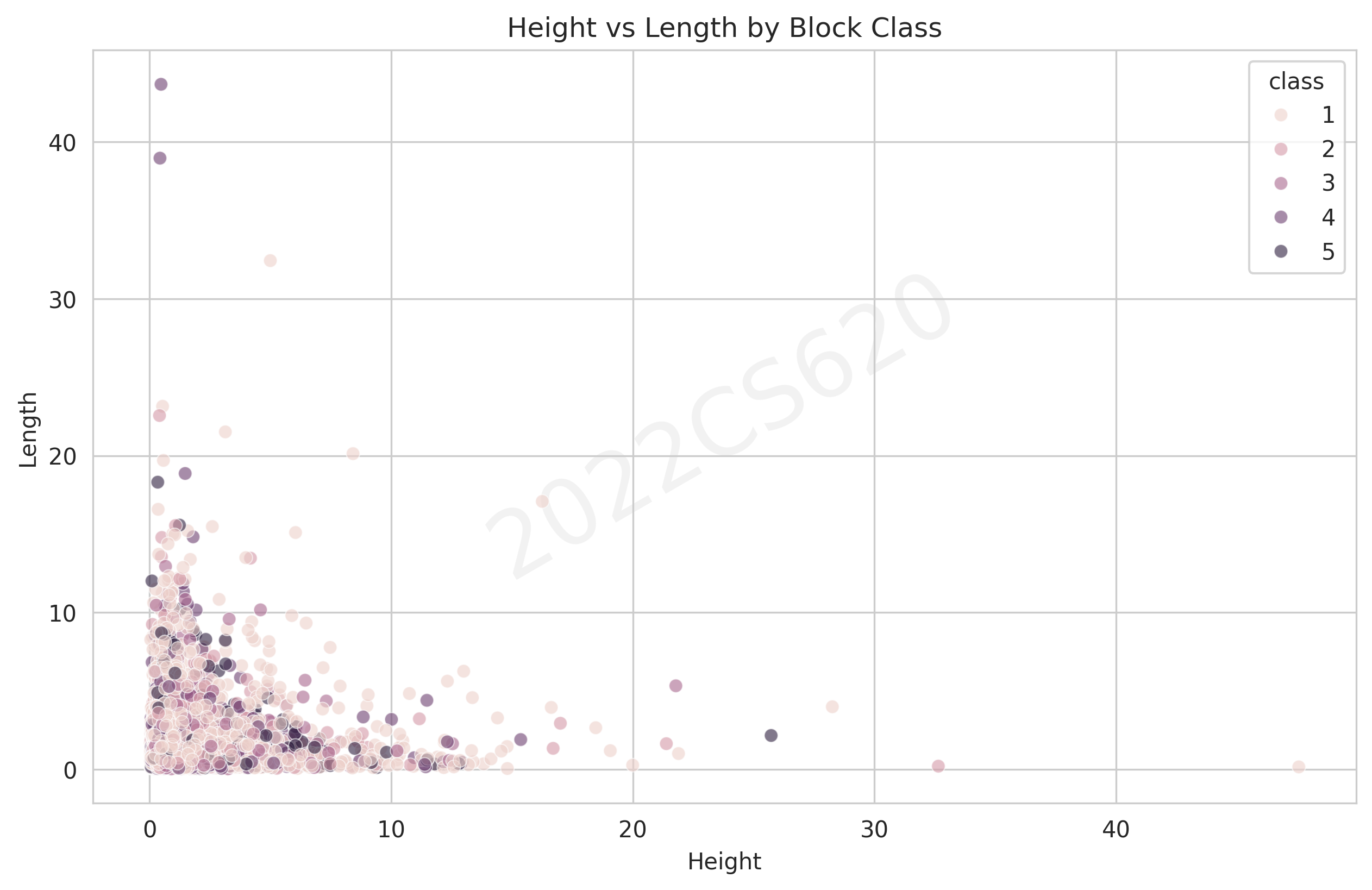


Figure 3: Relationship between block height and length, colored by class, showing distinct clustering patterns.

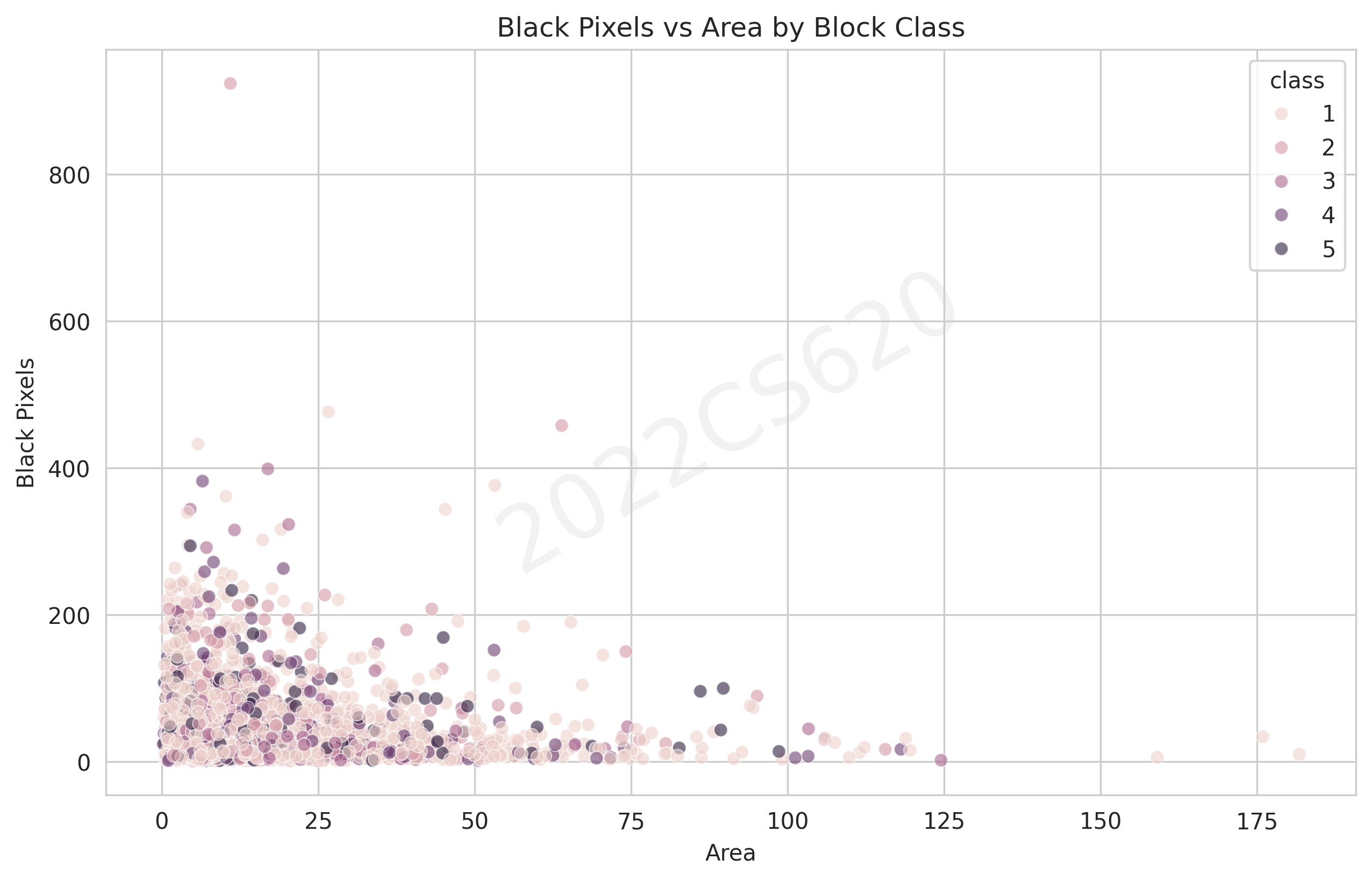


Figure 4: Relationship between block area and black pixels, revealing strong correlation and class-specific patterns.

# Multivariate Analysis

The multivariate analysis examines relationships between multiple features simultaneously to uncover complex patterns.

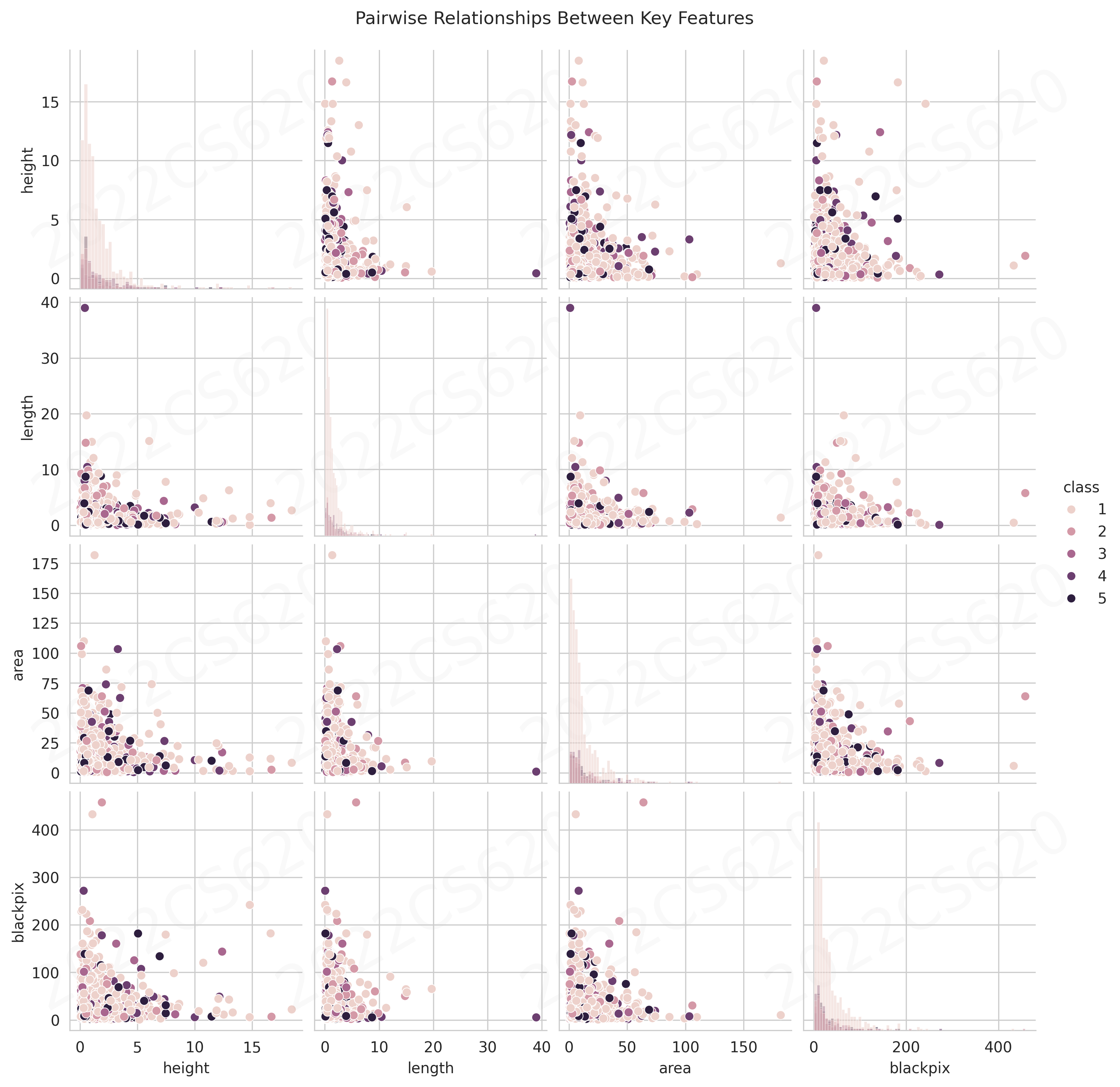


Figure 5: Pairwise relationships between key features showing complex interactions and class separation.

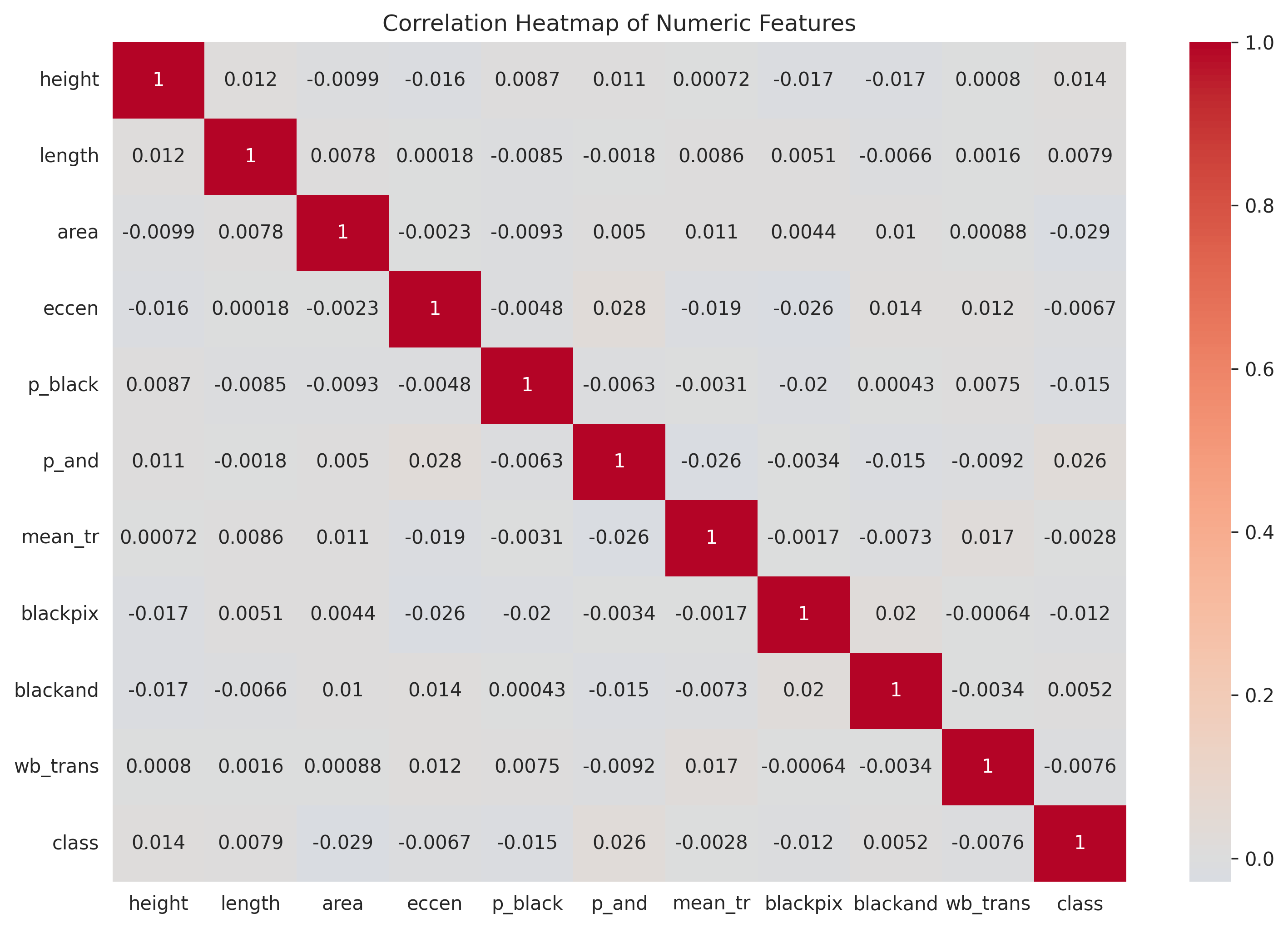


Figure 6: Correlation heatmap of numeric features highlighting strong relationships between certain features.

# Conclusions

The analysis reveals several important insights about the page blocks dataset:  
1. There is significant class imbalance with text blocks being the most common type.  
2. Block dimensions (height, length, area) show distinct patterns for different classes.  
3. Strong correlations exist between physical measurements and pixel-based features.  
4. The features provide good separation between different block classes, suggesting potential for accurate classification.