PAGE BLOCKS CLASSIFICATION

Presented by Shangzhi LOU,
Wenbo SUI, Limin TIAN

CONTENT

RESEARCH PURPOSES

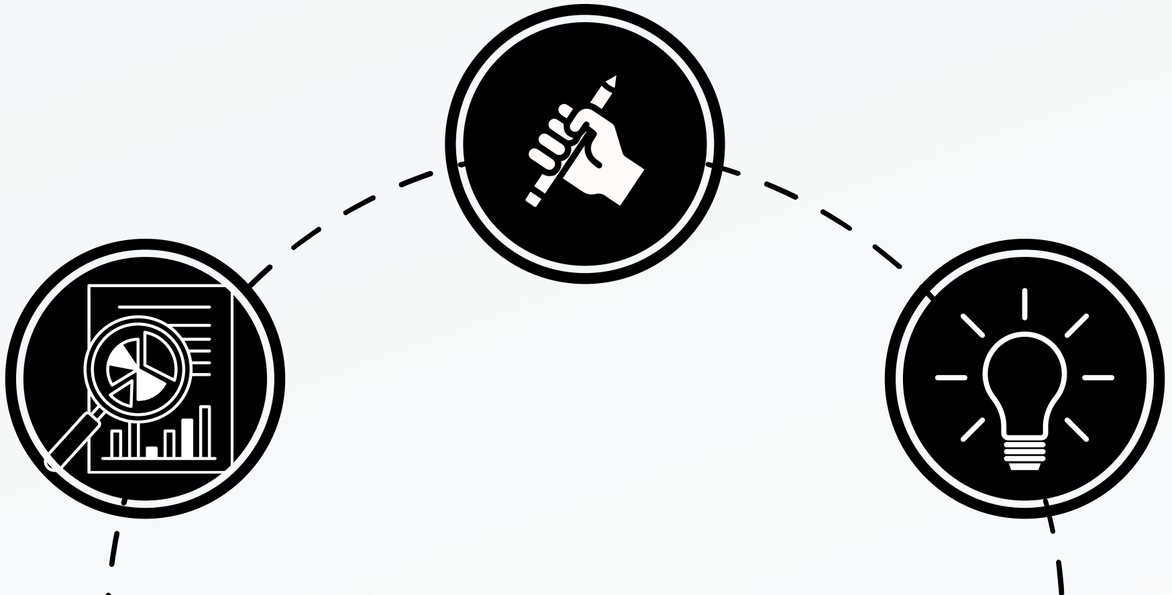
- **Information extraction and indexing:** Page layout analysis can help identify and extract different types of information blocks such as text, images, and tables.

- **Learning and AI applications:** Page block classification is a fundamental task in the field of machine vision and AI, helping to improve the ability of computer vision to identify document content.

- **Digital Archiving:** When digitizing paper documents, page chunking can help create a more structured digital copy. This is critical for long-term storage, information retrieval and digital protection.

EXPLORE DATASET

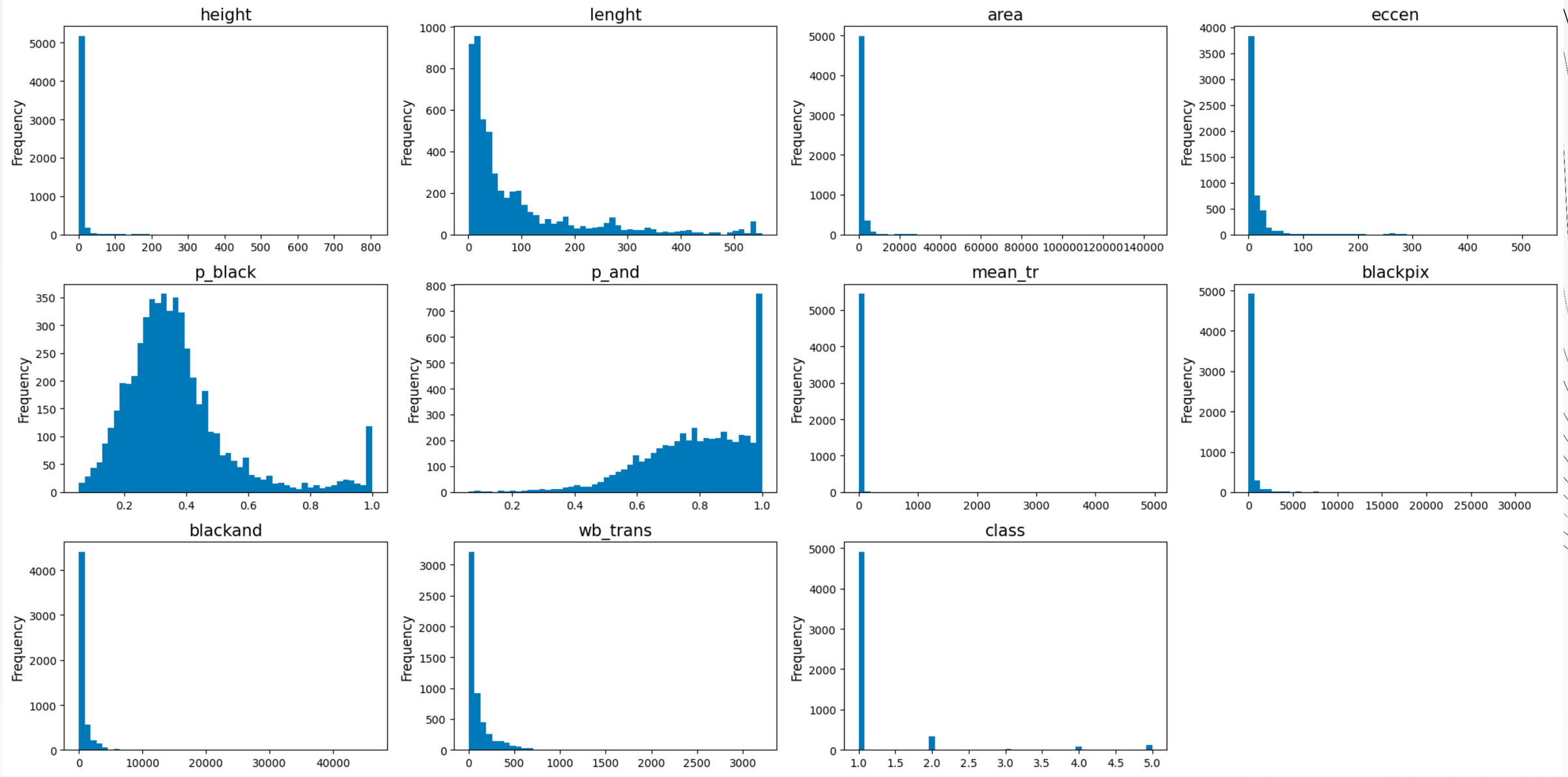
The description of the dataset → Dataset preprocessing → The observation of the dataset



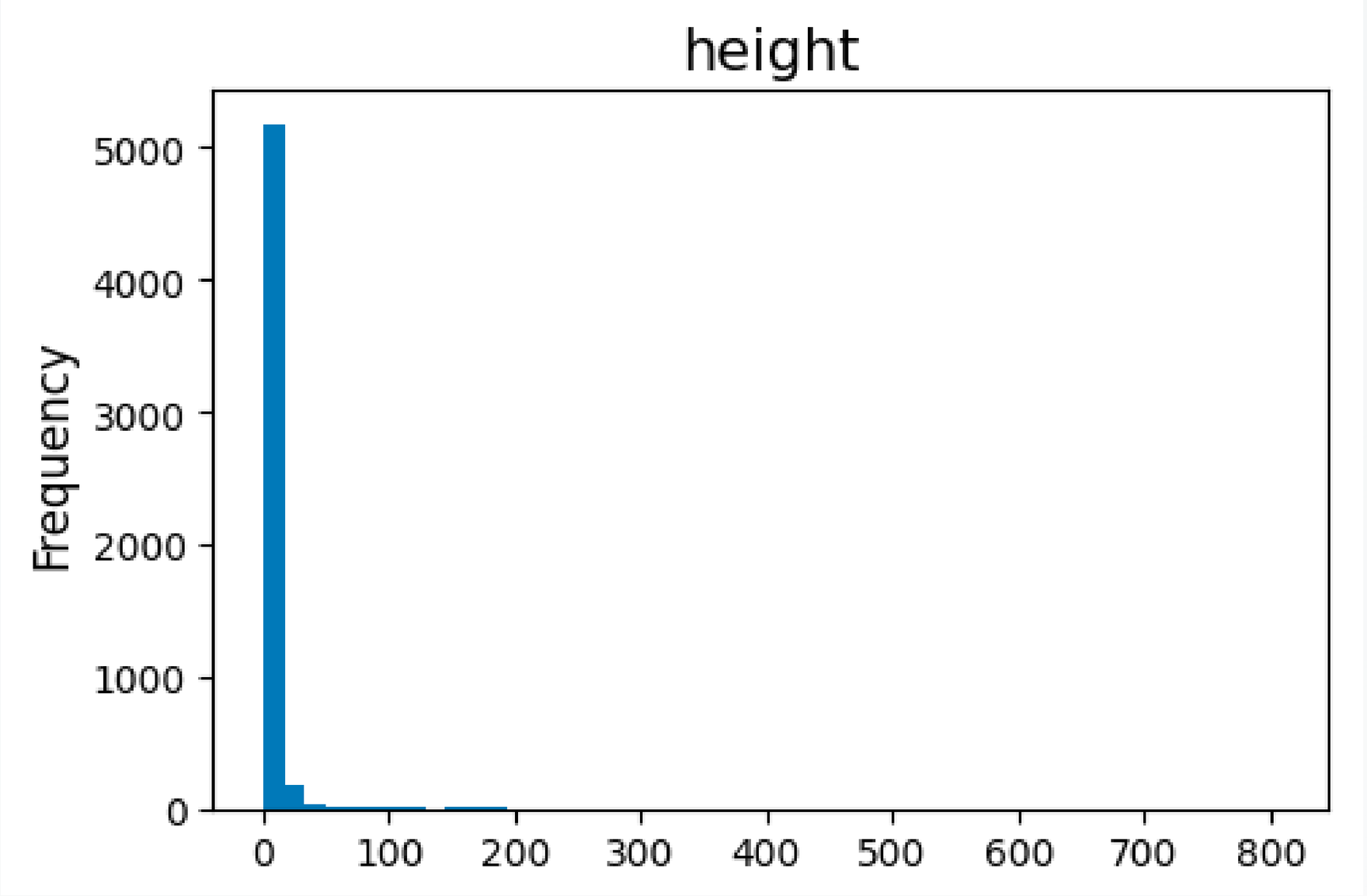
THE DESCRIPTION OF THE DATASET

- Sample size: **5473** samples.
- Number of features: There are **10** feature columns ('height', 'lenght', 'area', 'eccen', 'p_black', 'p_and', 'mean_tr', 'blackpix', 'blackand', 'wb_trans'), and 1 target column or category column ('class').
- Feature type: Some features are **integer types** ('height', 'lenght', 'area', 'blackpix', 'blackand', 'wb_trans'), and the rest are **continuous types** ('eccen', 'p_black', 'p_and', 'mean_tr').
- Target variable: "The 'class' column serves as the target variable, containing five classes: text (1), horizontal line (2), picture (3), vertical line (4), and graphic (5).

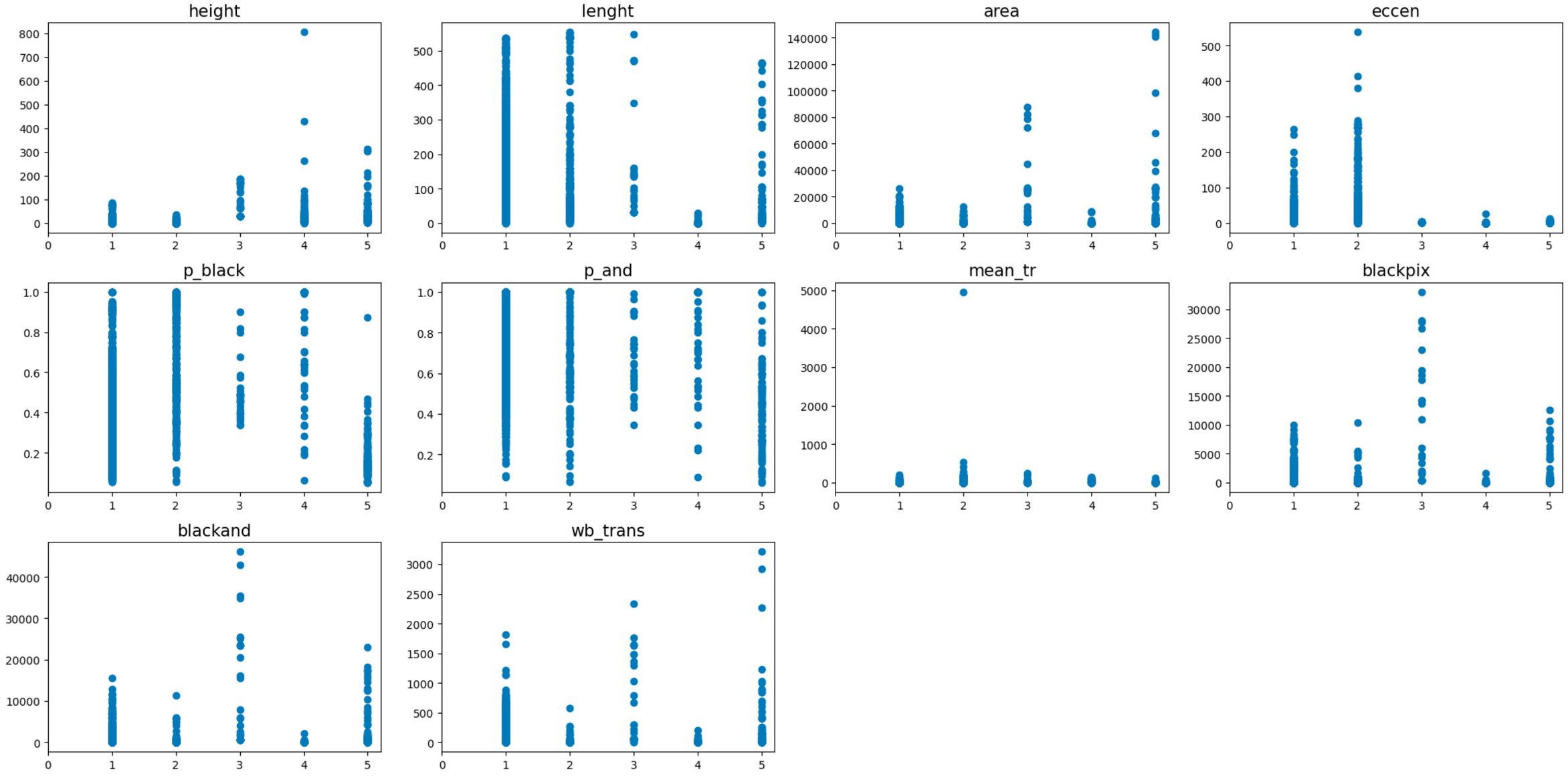
DISTRIBUTION OF ALL ATTRIBUTES



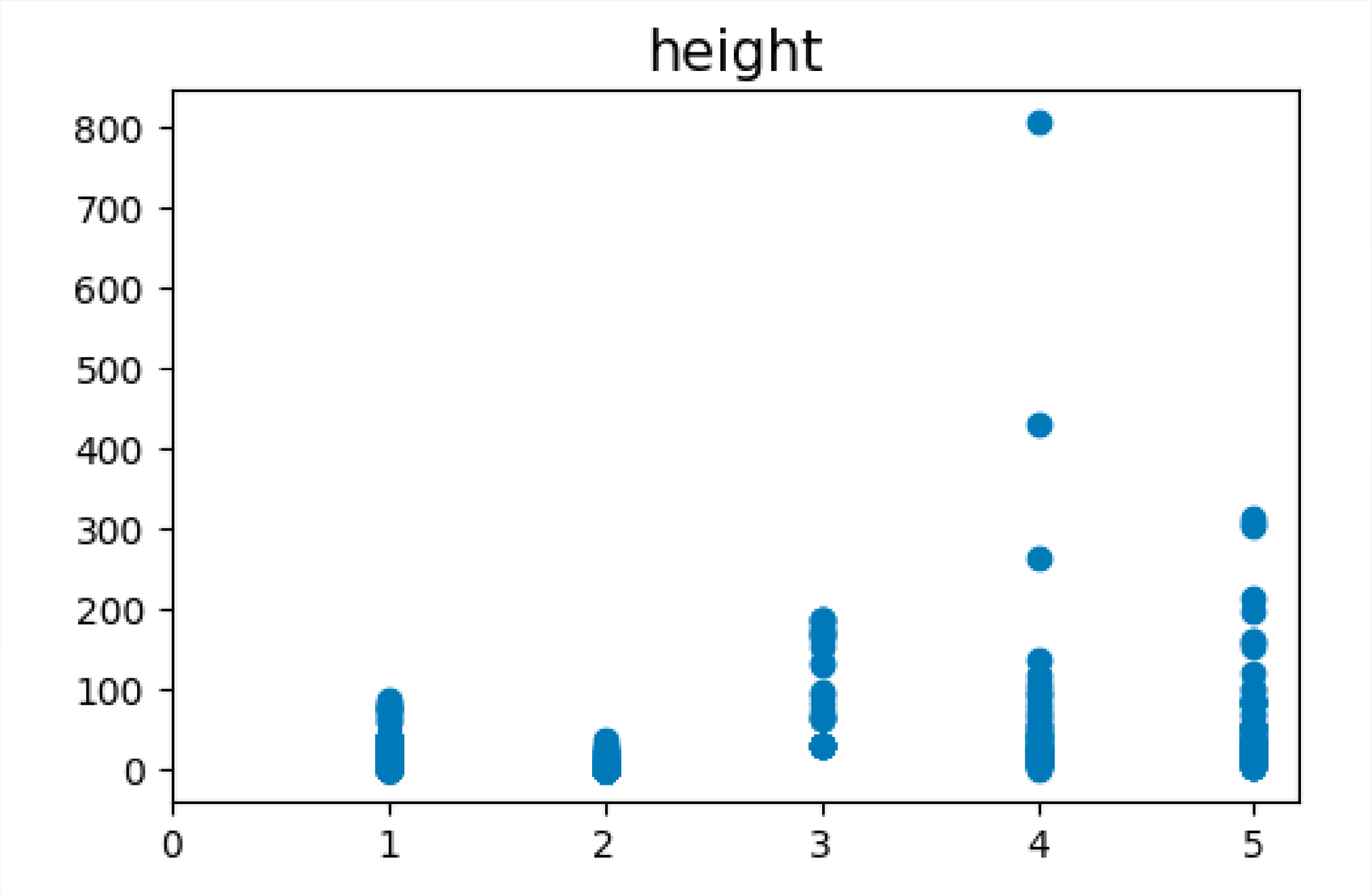
DISTRIBUTION OF ALL ATTRIBUTES



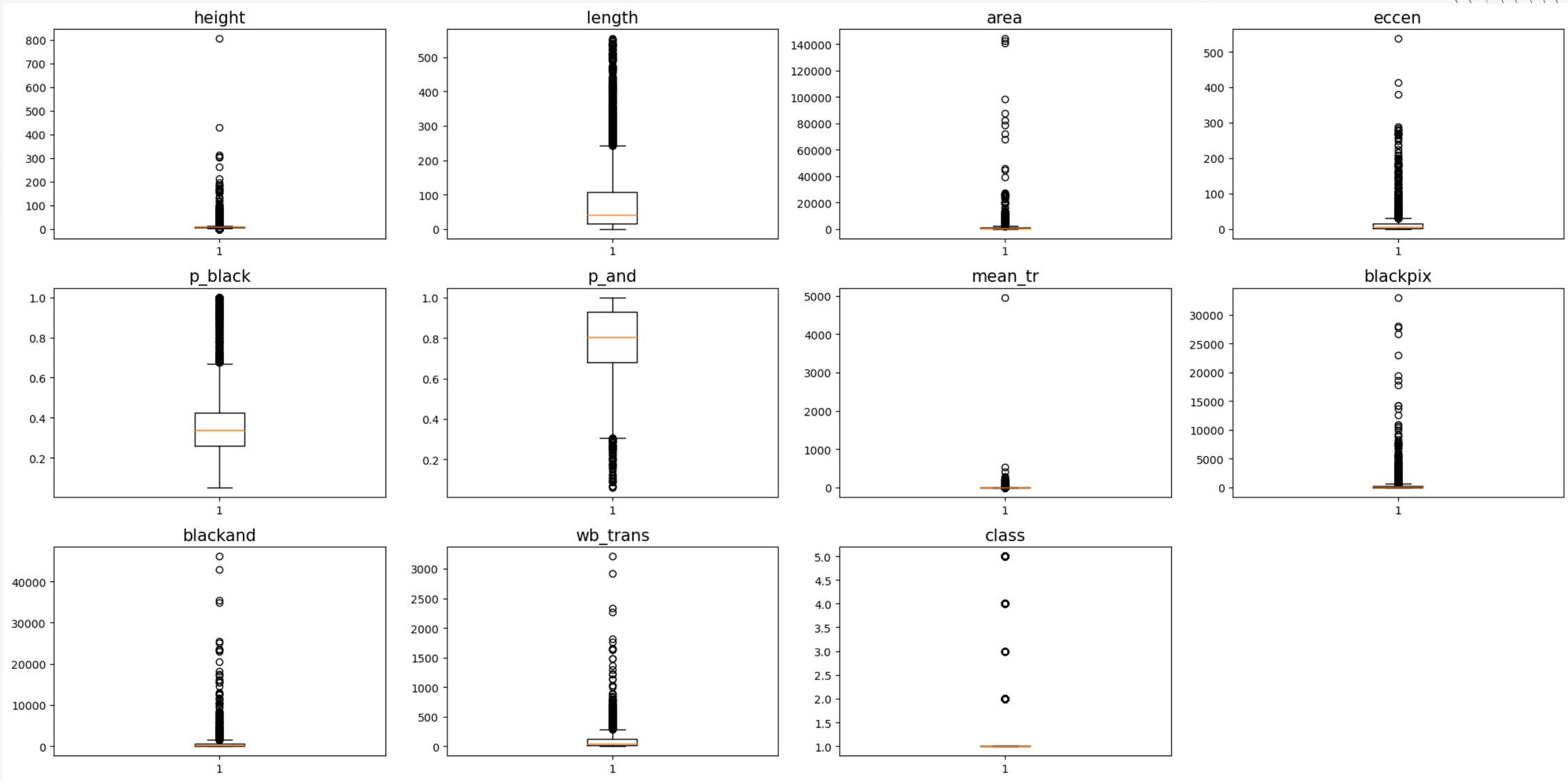
RELATIONSHIP BETWEEN OTHER ATTRIBUTES WITH TARGET ATTRIBUTE



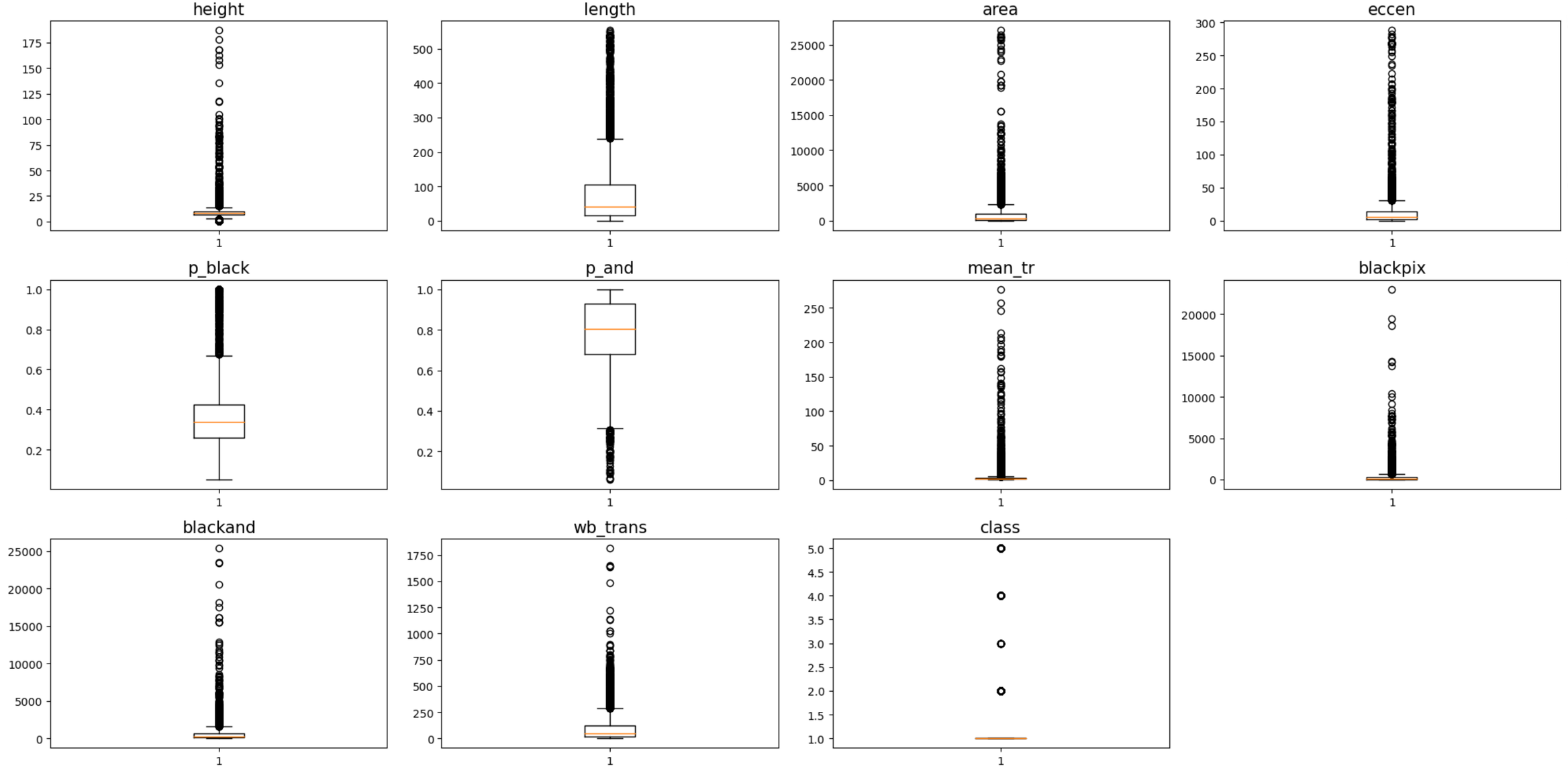
RELATIONSHIP BETWEEN OTHER ATTRIBUTES WITH TARGET ATTRIBUTE



BOXPLOT OF ALL ATTRIBUTES (OUTLIER DETECTION)

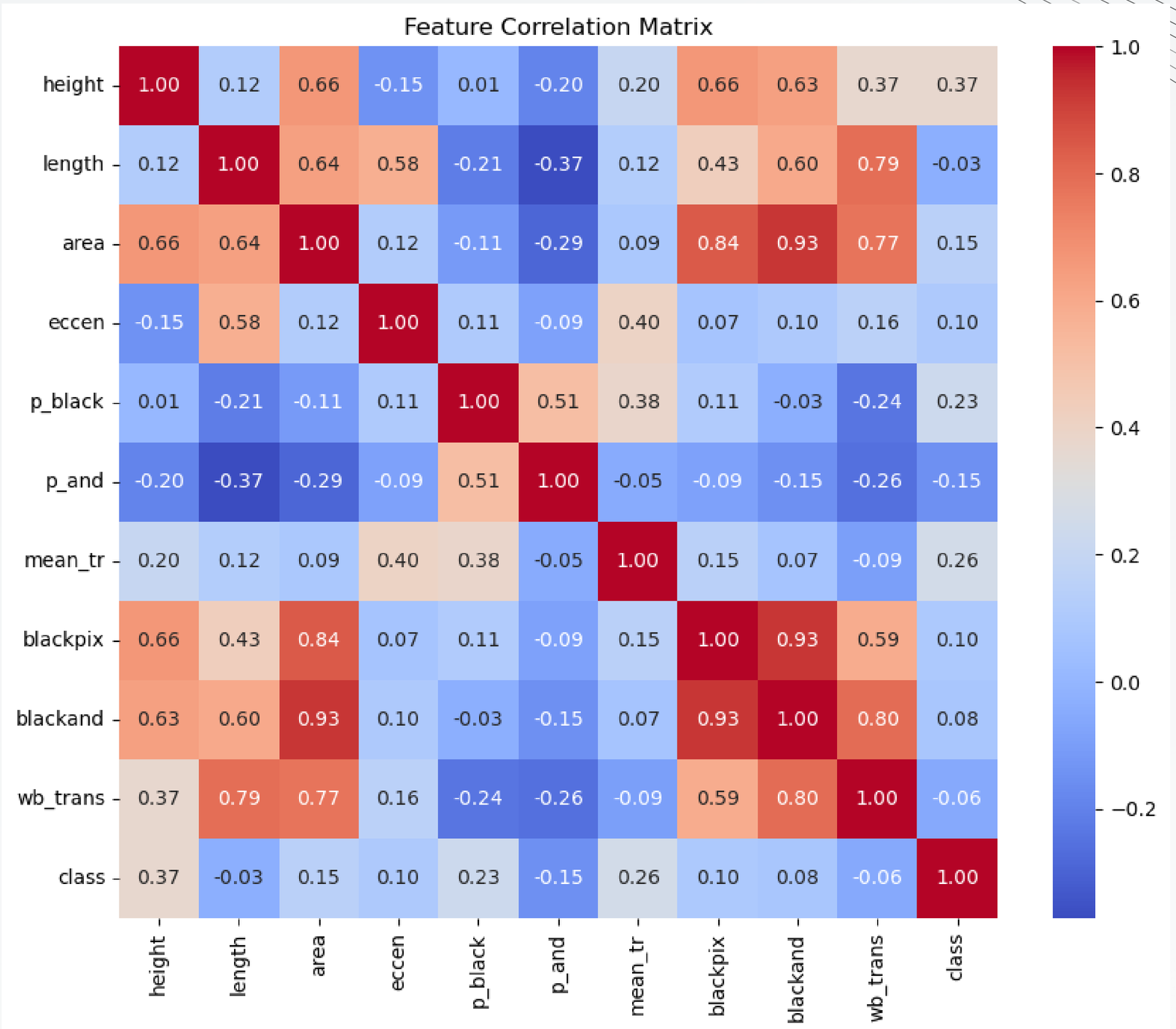


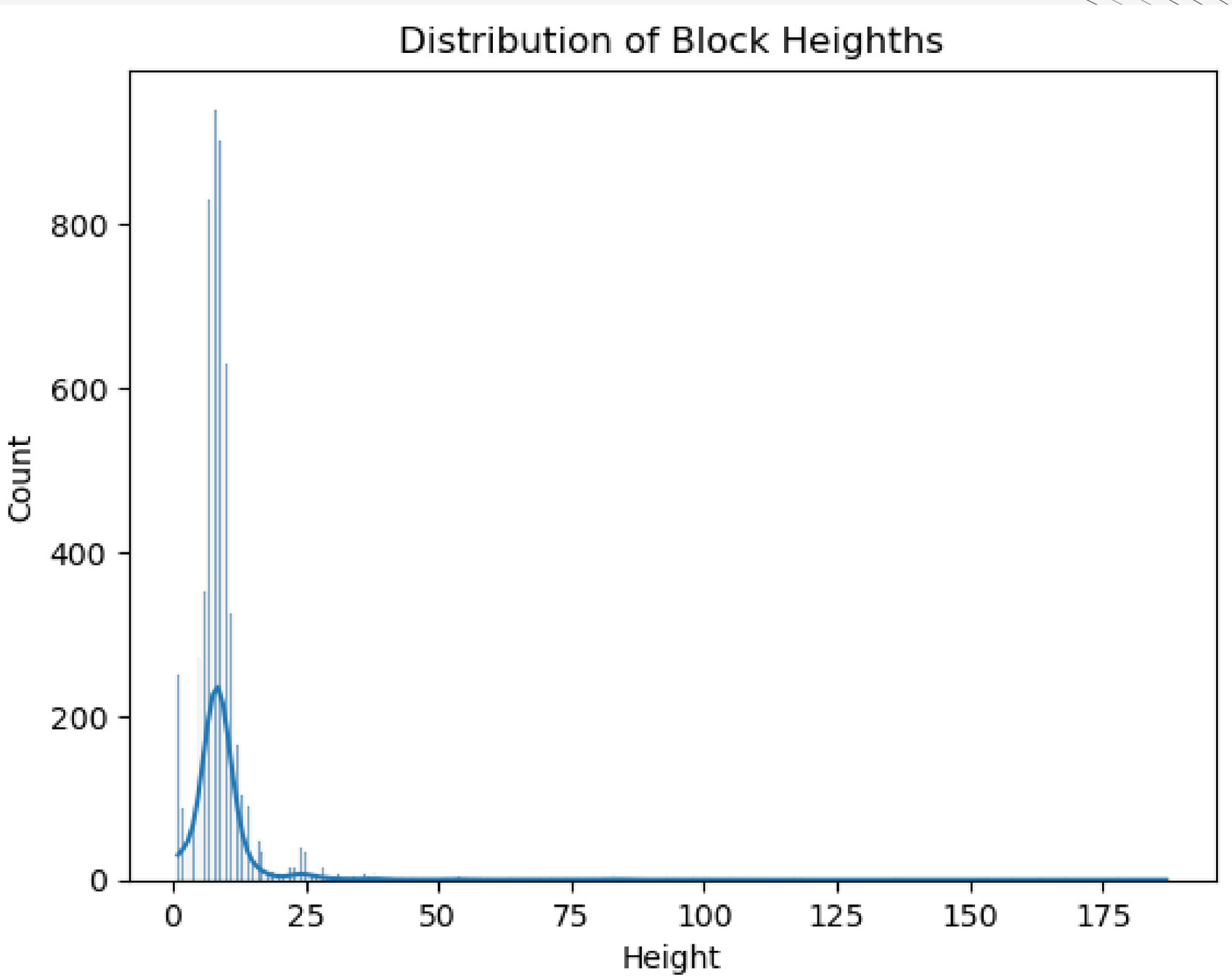
AFTER REMOVING OUTLIERS

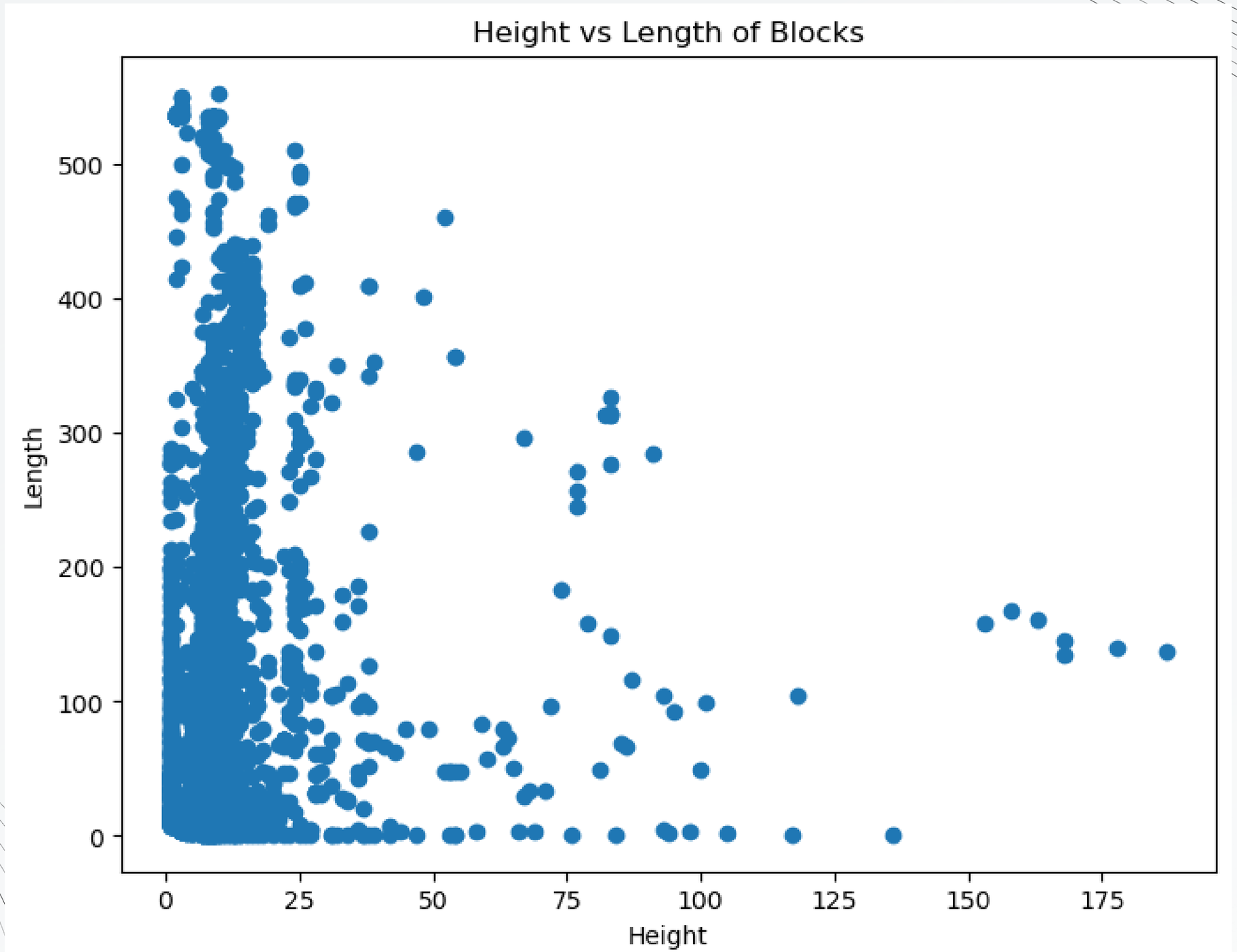


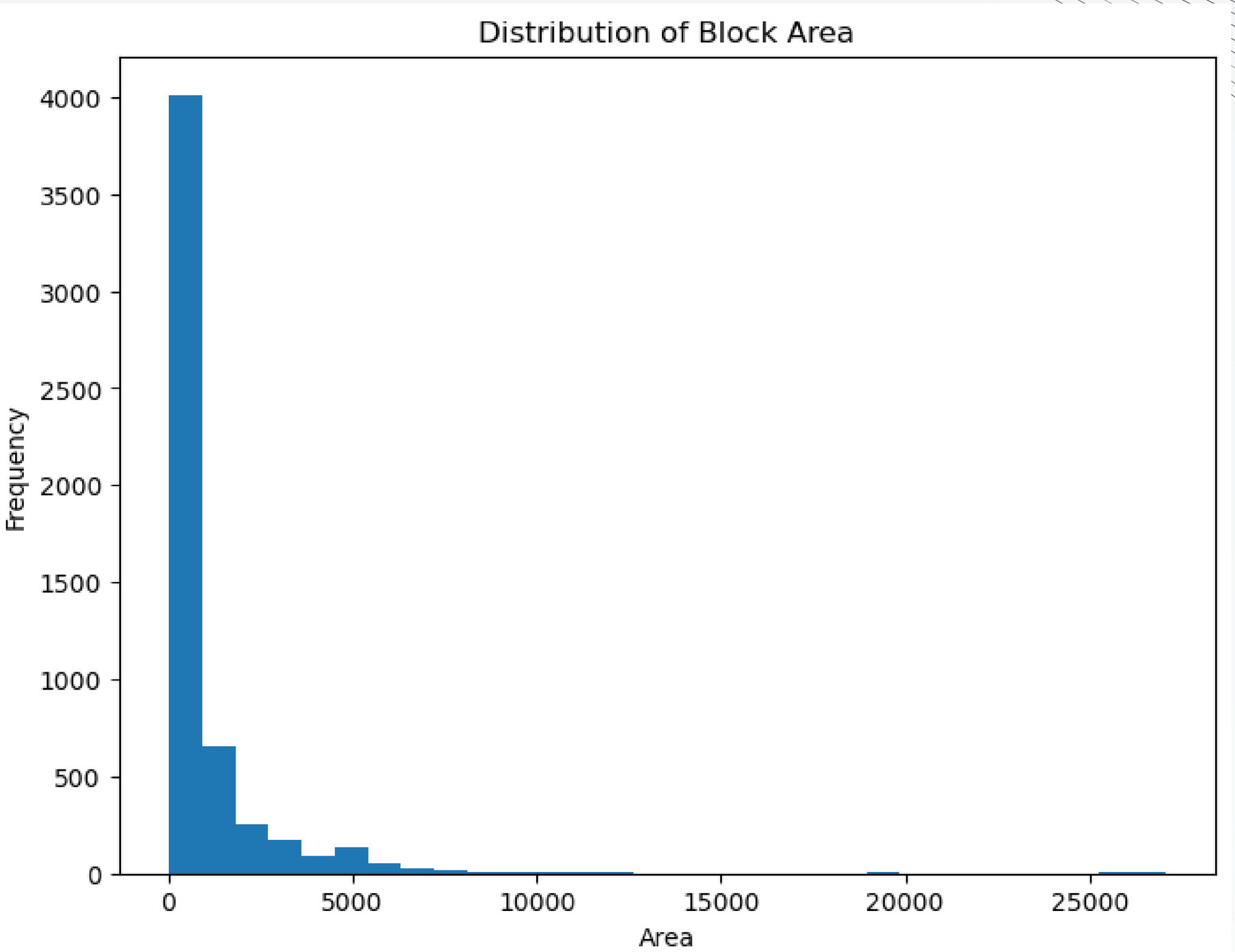
DATASET RESULTS

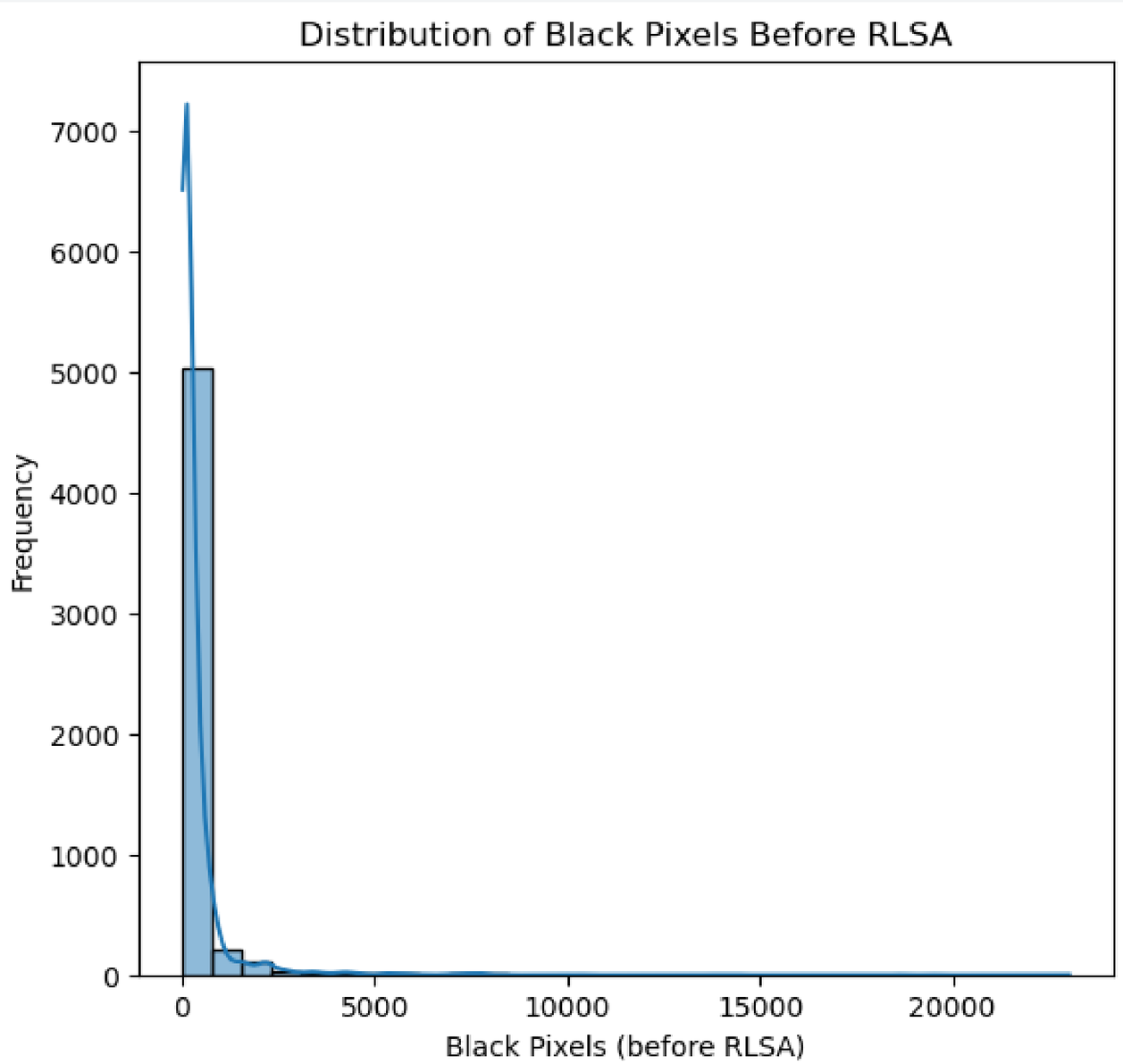
- Data cleaning: There are **no missing values** in the data set.
- Outlier Detection: By setting conditions, some **extreme values** are removed: ep, 'height' < 250, 'area' < 35000, to improve the accuracy of analysis and the generalization ability of the model.
- Conversion and encoding: No conversion or encoding is necessary in this simulation because all data is **numeric**.
- Normalization: All features except the target variable are **standardized**, that is, the mean is subtracted and then divided by the standard deviation, so that the mean of the feature becomes 0 and the standard deviation becomes 1.

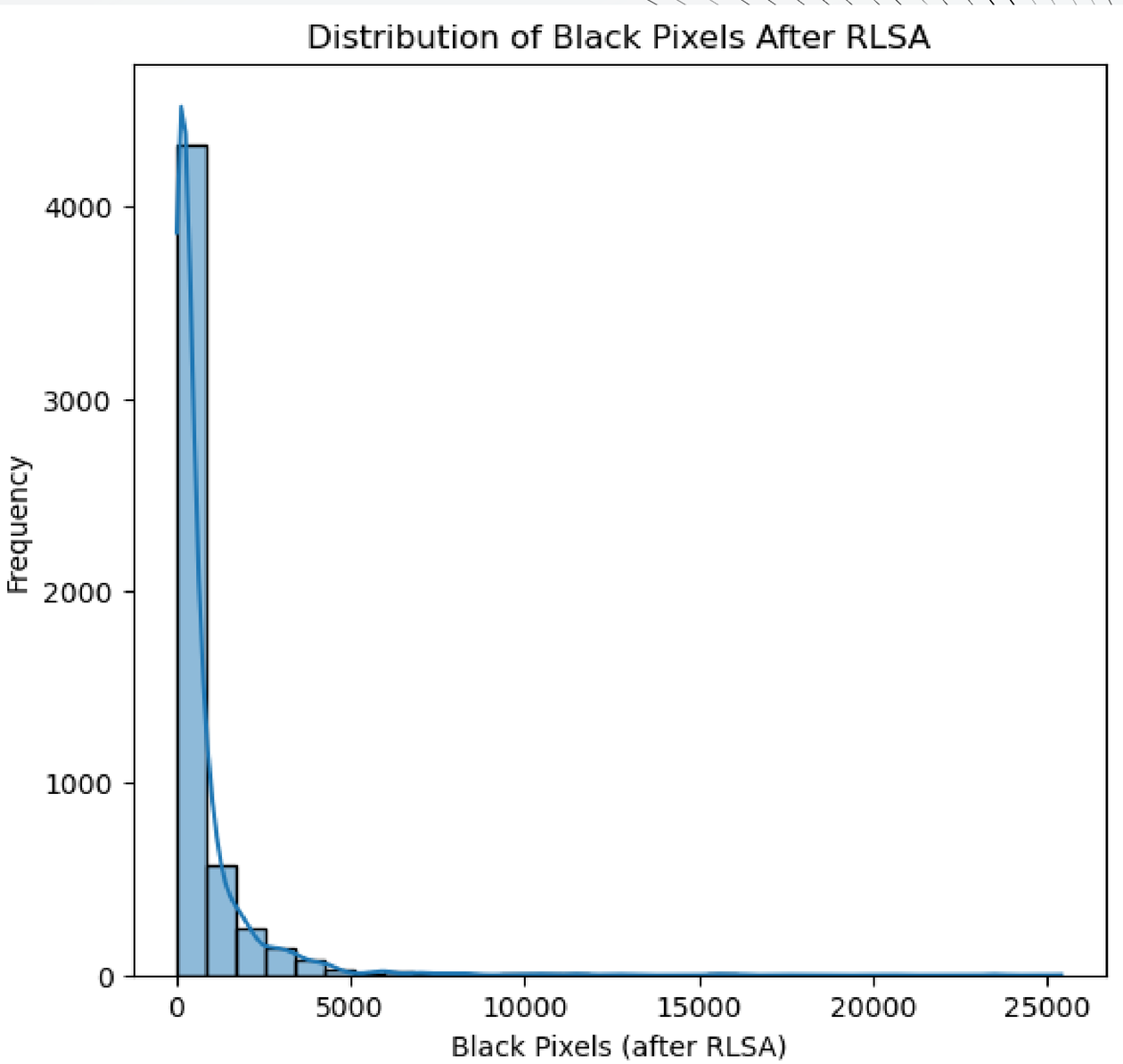

Feature Correlation Matrix
height 1.00 0.12 0.66 -0.15 0.01 -0.20 0.20 0.66 0.63 0.37 0.37
length 0.12 1.00 0.64 0.58 -0.21 -0.37 0.12 0.43 0.60 0.79 -0.03
area 0.66 0.64 1.00 0.12 -0.11 -0.29 0.09 0.84 0.93 0.77 0.15
eccen -0.15 0.58 0.12 1.00 0.11 -0.09 0.40 0.07 0.10 0.16 0.10
p_black 0.01 -0.21 -0.11 0.11 1.00 0.51 0.38 0.11 -0.03 -0.24 0.23
p_and -0.20 -0.37 -0.29 -0.09 0.51 1.00 -0.05 -0.09 -0.15 -0.26 -0.15
mean_tr 0.20 0.12 0.09 0.40 0.38 -0.05 1.00 0.15 0.07 -0.09 0.26
blackpix 0.66 0.43 0.84 0.07 0.11 -0.09 0.15 1.00 0.93 0.59 0.10
blackand 0.63 0.60 0.93 0.10 -0.03 -0.15 0.07 0.93 1.00 0.80 0.08
wb_trans 0.37 0.79 0.77 0.16 -0.24 -0.26 -0.09 0.59 0.80 1.00 -0.06
class 0.37 -0.03 0.15 0.10 0.23 -0.15 0.26 0.10 0.08 -0.06 1.00
height length area eccen p_black p_and mean_tr blackpix blackand wb_trans class
1.0 0.8 0.6 0.4 0.2 0.0 −0.2


Distribution of Block Heighths
Count
800
600
400
200
0
0
25
50
75
100
125
150
175
Height

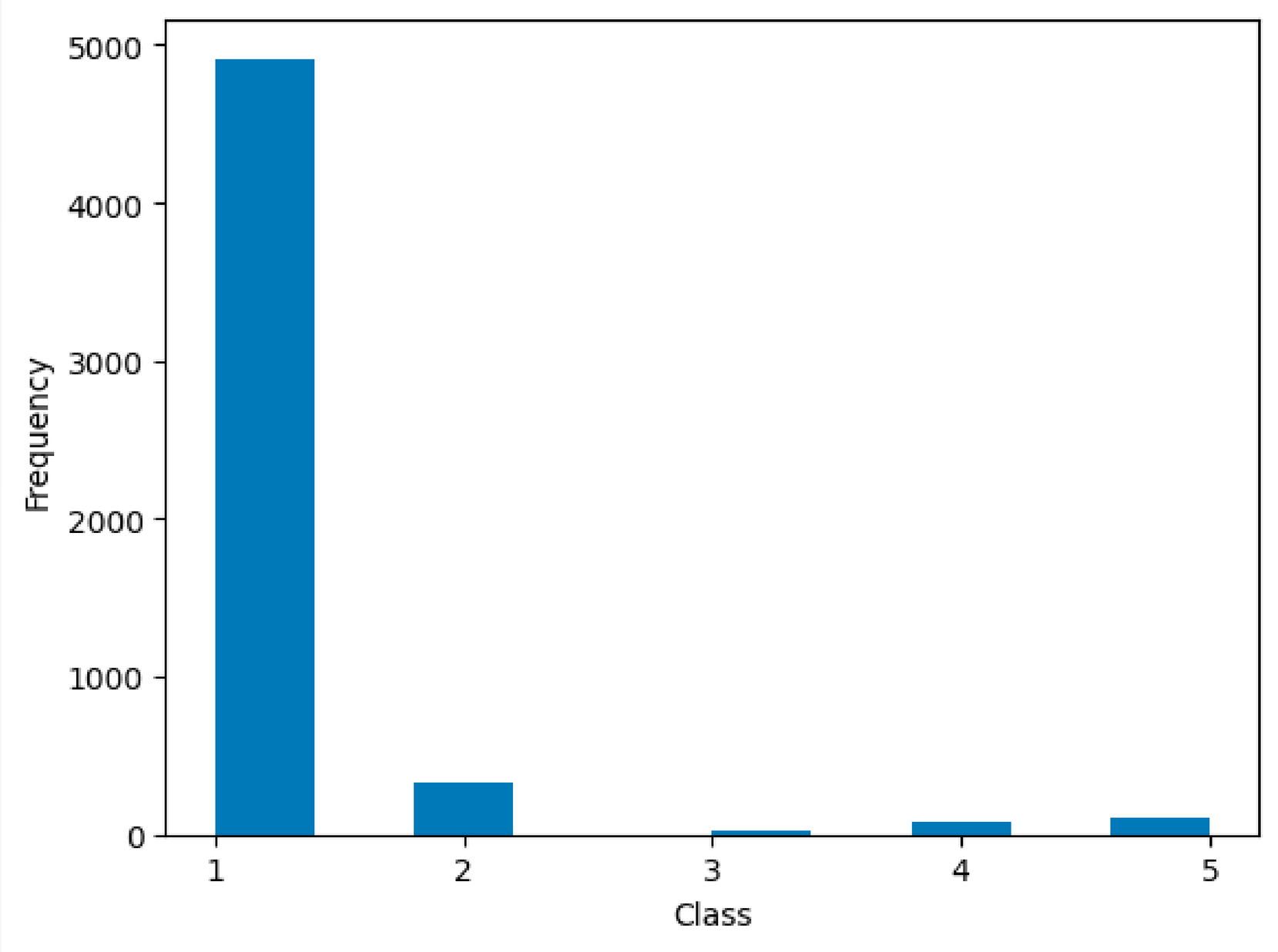

Height vs Length of Blocks
Length
Height
0
100
200
300
400
500
0
25
50
75
100
125
150
175


Distribution of Block Area
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
0
5000
10000
15000
20000
25000
Area


Distribution of Black Pixels Before RLSA
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
Black Pixels (before RLSA)


Distribution of Black Pixels After RLSA
Frequency
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
Black Pixels (after RLSA)

VISUALIZATIONS



OBSERVATION OF THE DATASET

- The dataset is highly imbalanced, most of examples belong to class 1.
- Most blocks have a small height, suggests that blocks of smaller height are more common.
- The 'wb_trans' has a high correlation with 'blackpix' and 'blackand'.
- The area of the block has a high correlation with height, length,blackpix, and blackand.
- The blackpix has a high correlation with blackand.

BASELINE MODEL OF LOGISTIC REGRESSION

The dataset involves **continuous and integer** features, and the goal is classification.

For binary classification problems, this is a good starting point because it is simple and easy to implement, while the output probabilities provide a clear classification decision.

RANDOM FOREST MODEL

Random Forest is an ensemble learning method that improves the performance and robustness of the overall model by combining multiple decision trees.

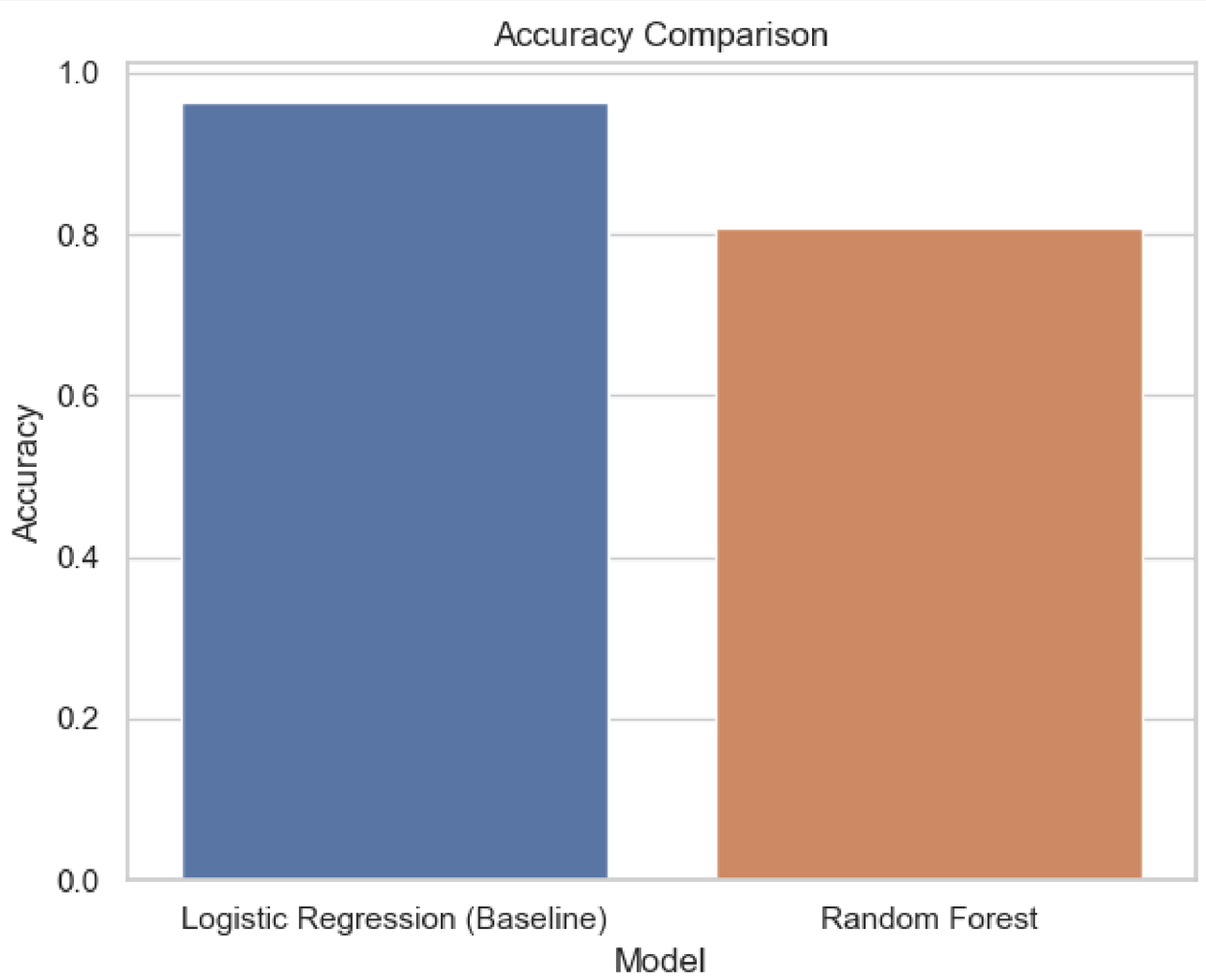
(1) **High performance:** Random forests generally perform well on various types of data sets. It is very effective for processing high-dimensional, non-linear and complex relationship data sets.

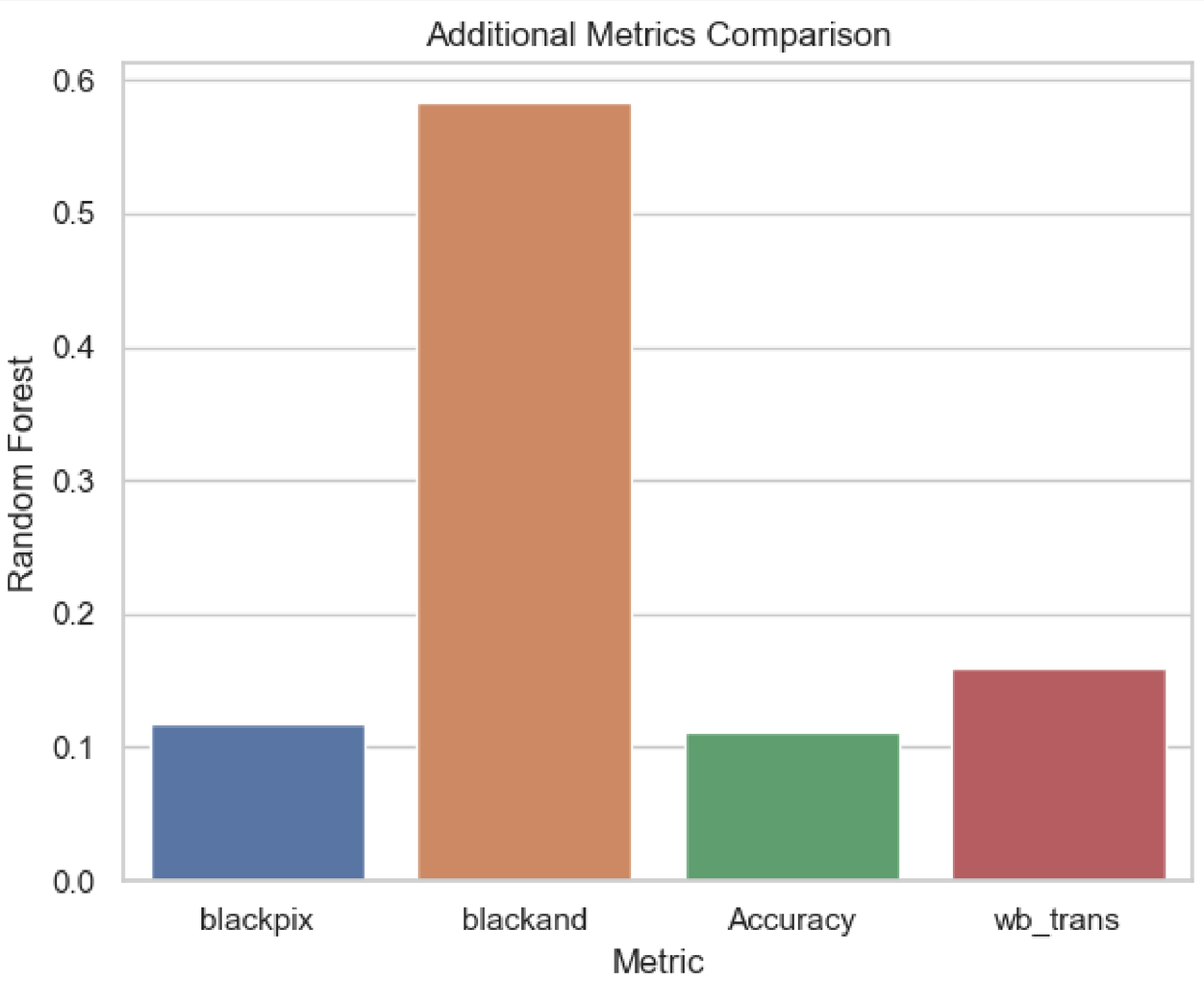
(2) **Feature importance:** Random forests provide an intuitive way to evaluate feature importance. By observing how each feature is used across multiple decision trees, we can determine which features are critical to the model's predictions.

(3) Handle imbalanced data: Random forests can handle imbalanced data sets and can balance predictions between classes through a voting mechanism.

(1) PREDICTING A SINGLE TARGET (REGRESSION PROBLEM)

```
Accuracy: 0.8065994500458296
Accuracy (Random Forest): 0.8065994500458296
Feature Importance:
eccen: 0.2413
area: 0.1598
length: 0.0962
blackpix: 0.0905
blackand: 0.0879
mean_tr: 0.0876
p_black: 0.0828
p_and: 0.0791
wb_trans: 0.0579
class: 0.0168
```


Accuracy Comparison
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Logistic Regression (Baseline)
Random Forest
Model


Additional Metrics Comparison
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Random Forest
blackpix
blackand
Accuracy
wb_trans
Metric

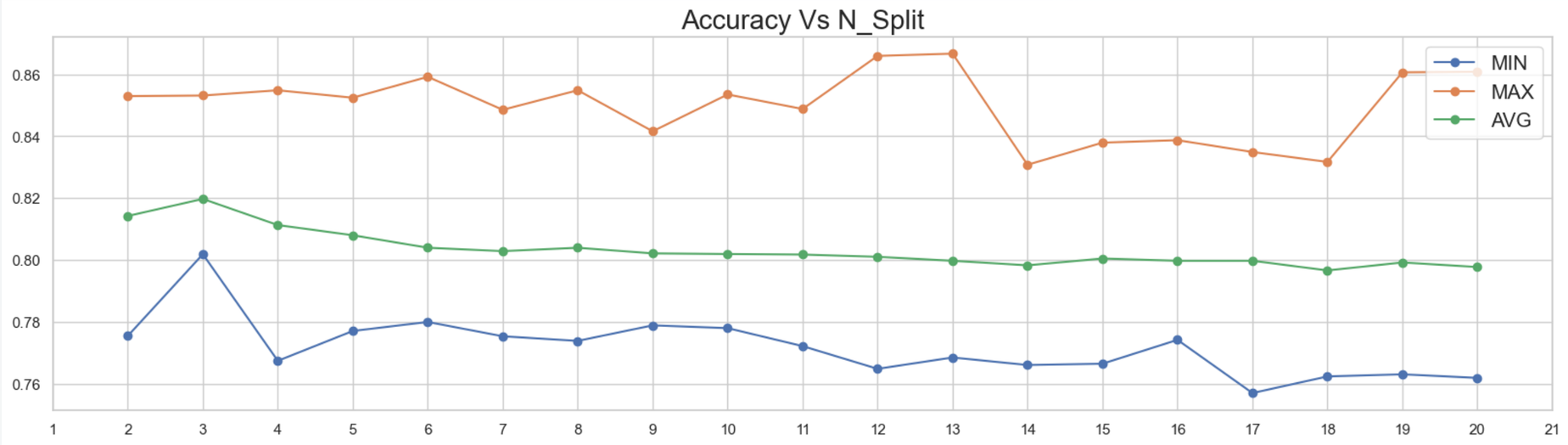
(2) MULTI-OBJECTIVE REGRESSION PROBLEM

```
Mean Squared Error: 82160.83161707914
```

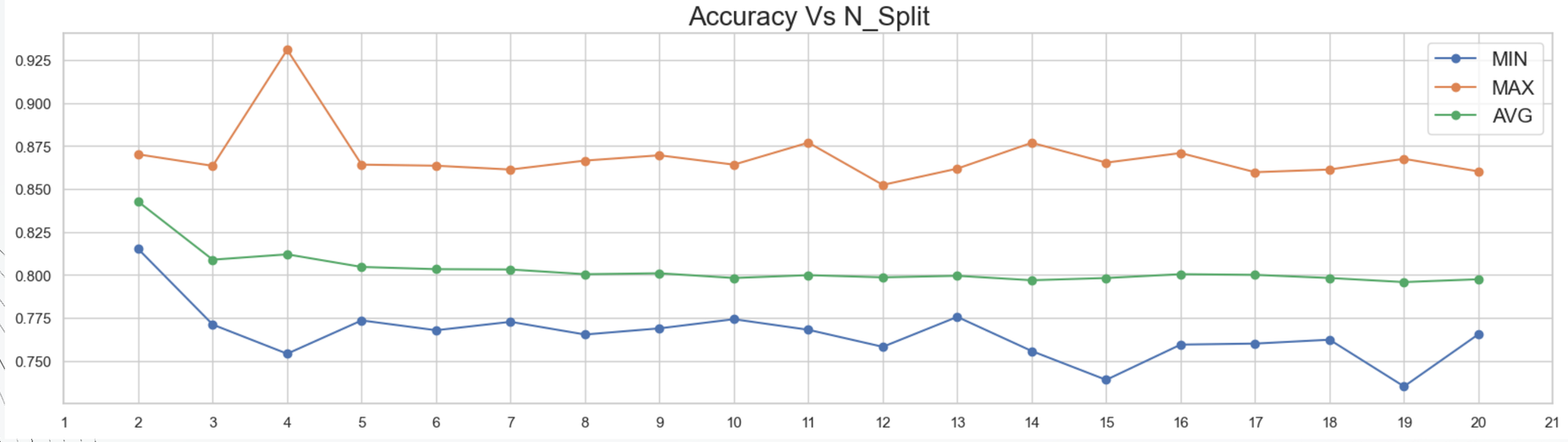
NOt good...

MODEL EVALUATION

CALCULATE THE CROSS VALIDATION SCORE ON DIFFERENT SPLIT POINT IN K-FOLD



K-FOLD



STRATIFIED K-FOLD

RESULTS AND DISCUSSION

- Using GaussianNB, the model achieves an average accuracy of approximately 88%.
- In the 5-fold and 10-fold cross-validation, the interval between the minimum and maximum accuracy values is not large, indicating that the model has a certain stability.
- Try to create new features or transform existing features to improve the performance of the model. Handle imbalanced data sets through oversampling or undersampling.

REFERENCE

- Dataset:

https://archive.ics.uci.edu/dataset/78/page+blocks+classification

- Article:

1. Fushiki, Tadayoshi. "Estimation of prediction error by using K-fold cross-validation." Statistics and Computing 21 (2011): 137-146.
2. Anguita, Davide, et al. "The'K'in K-fold Cross Validation." ESANN. 2012.

THANK'S FOR WATCHING