# Wheat Kernel Categorization Using a Neural Network

## Introduction and Credits

In this project, we will attempt to classify three wheat varieties based on geometric attributes of their kernels. The dataset we use for training and testing was prepared by M. Charytanowicz et al. at the Institute of Mathematics and Computer Science of the John Paul II Catholic University in Lublin, Poland. It is available at the UCI Machine Learning Repository.

## Dataset Description

Using a soft X-ray technique, Charytanowicz et al. measured the following kernel features:

- Area $A$
- Perimeter $P$
- Compactness
- Length
- Width
- Asymmetry coefficient
- Length of kernel groove

The kernels belong to the Kama, Rosa, and Canadian varieties, labeled in the dataset as categories 1, 2, and 3.

## Preliminary Processing

The numerical range of different features varies significantly. We noted, for instance, that while the area varies between 10.59 and 21.18 units, the compactness coefficient varies between just 0.8081 and 0.9183 units.

| | Area | Perimeter | Compactness | Length | Width | Asymmetry | Groove Length |
|---|---|---|---|---|---|---|---|
| Min | 10.59 | 12.41 | 0.8081 | 4.899 | 2.630 | 0.7651 | 4.519 |
| Max | 21.18 | 17.25 | 0.9183 | 6.675 | 4.033 | 8.4560 | 6.550 |

In our first attempt with preprocessing, we scaled the range of each feature to be between the numbers 1 and 2. This prevents features with larger ranges from being represented disproportionally.

## Training and accuracy measures

We use a randomly selected 80/20 training/testing split. Thus 168 entries are used for training. Given the relatively small size of the training set, we run the training 5 times on different training/testing selections, and take the median loss and accuracy scores across runs. We found this step necessary as the loss and accuracy scores varied by more than 5% at times, making accessing both the absolute and relative performance of different training parameters difficult.

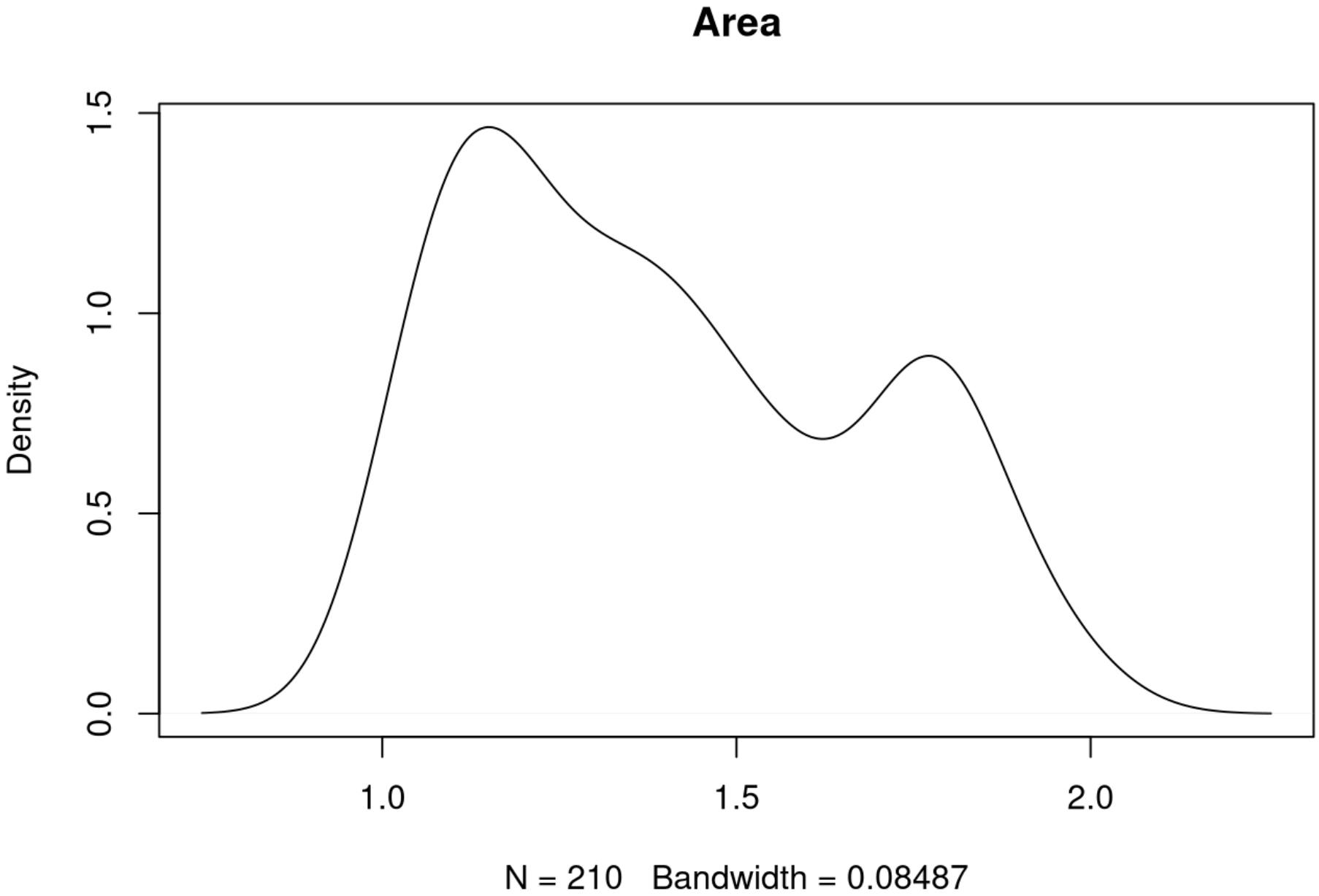
## Feature Selection: First Round

### Correlations

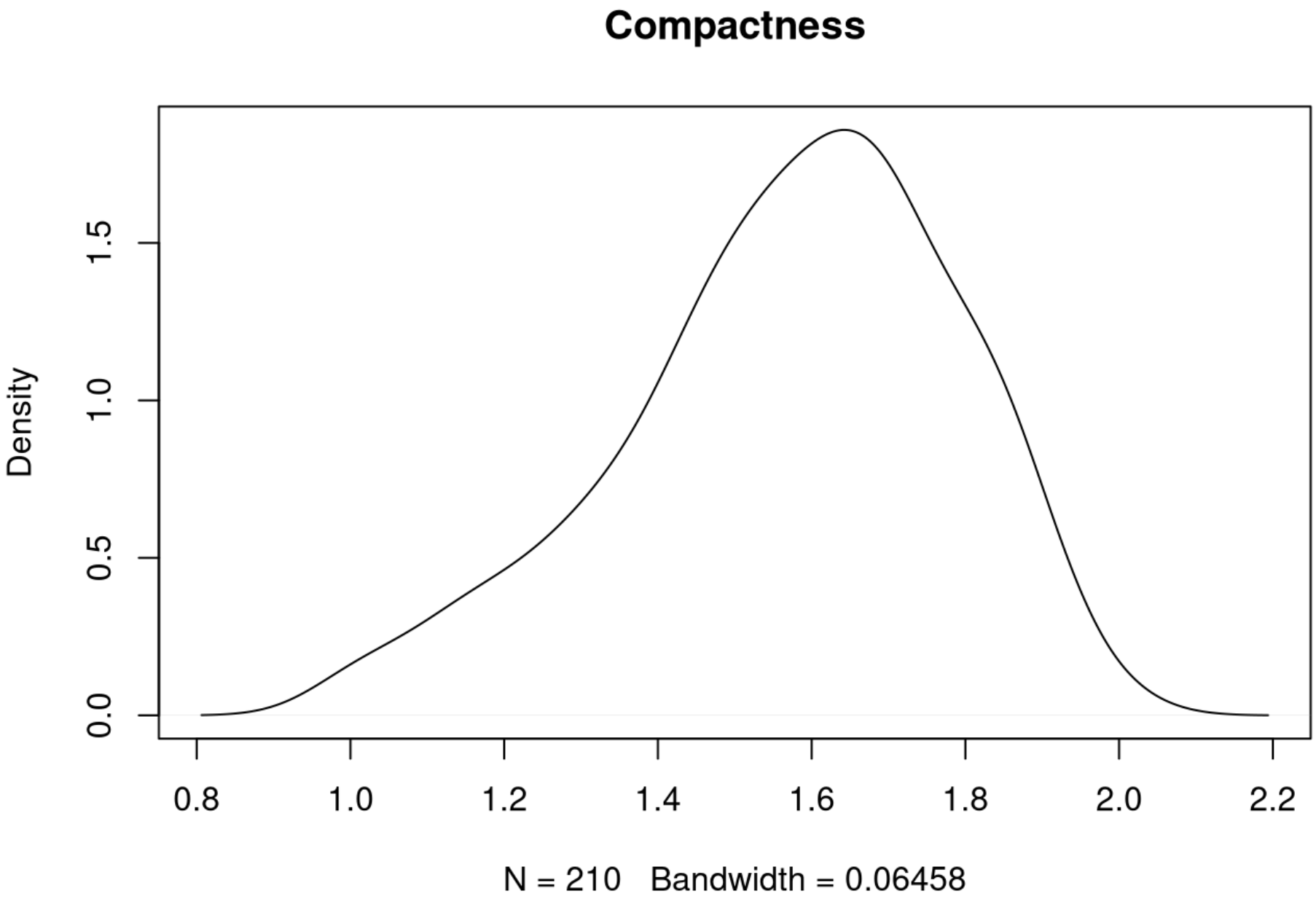
We use the following Pearson correlation matrix to aid with feature selection. Among highly correlated features, we are interested in selecting features whose kernel density plots (seen below) most clearly suggest multiple distributions exist (the hope being that each of these distributions is associated with a category).

| | Area | Perimeter | Compactness | Length | Width | Asymmetry | Groove Length | Variety |
|---|---|---|---|---|---|---|---|---|
| Area | 1 | 0.994 | 0.608 | 0.95 | 0.971 | -0.23 | 0.864 | -0.346 |
| Perimeter | 0.994 | 1 | 0.529 | 0.972 | 0.945 | -0.217 | 0.891 | -0.328 |
| Compactness | 0.608 | 0.529 | 1 | 0.368 | 0.762 | -0.331 | 0.227 | -0.531 |
| Length | 0.95 | 0.972 | 0.368 | 1 | 0.86 | -0.172 | 0.933 | -0.257 |
| Width | 0.971 | 0.945 | 0.762 | 0.86 | 1 | -0.258 | 0.749 | -0.423 |
| Asymmetry | -0.23 | -0.217 | -0.331 | -0.172 | -0.258 | 1 | -0.011 | 0.577 |
| Groove Length | 0.864 | 0.891 | 0.227 | 0.933 | 0.749 | -0.011 | 1 | 0.024 |
| Variety | -0.346 | -0.328 | -0.531 | -0.257 | -0.423 | 0.577 | 0.024 | 1 |

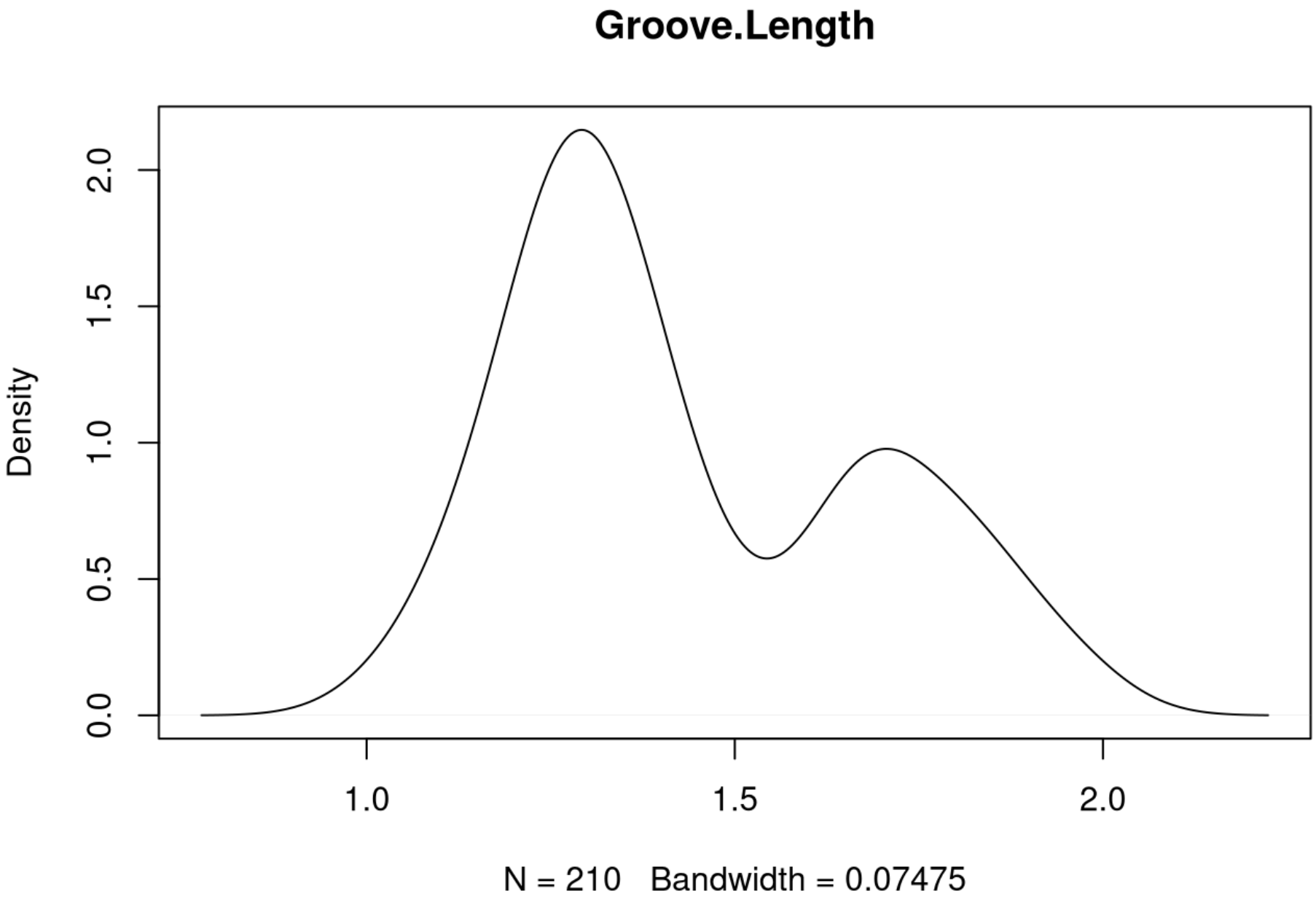
**Distributions**



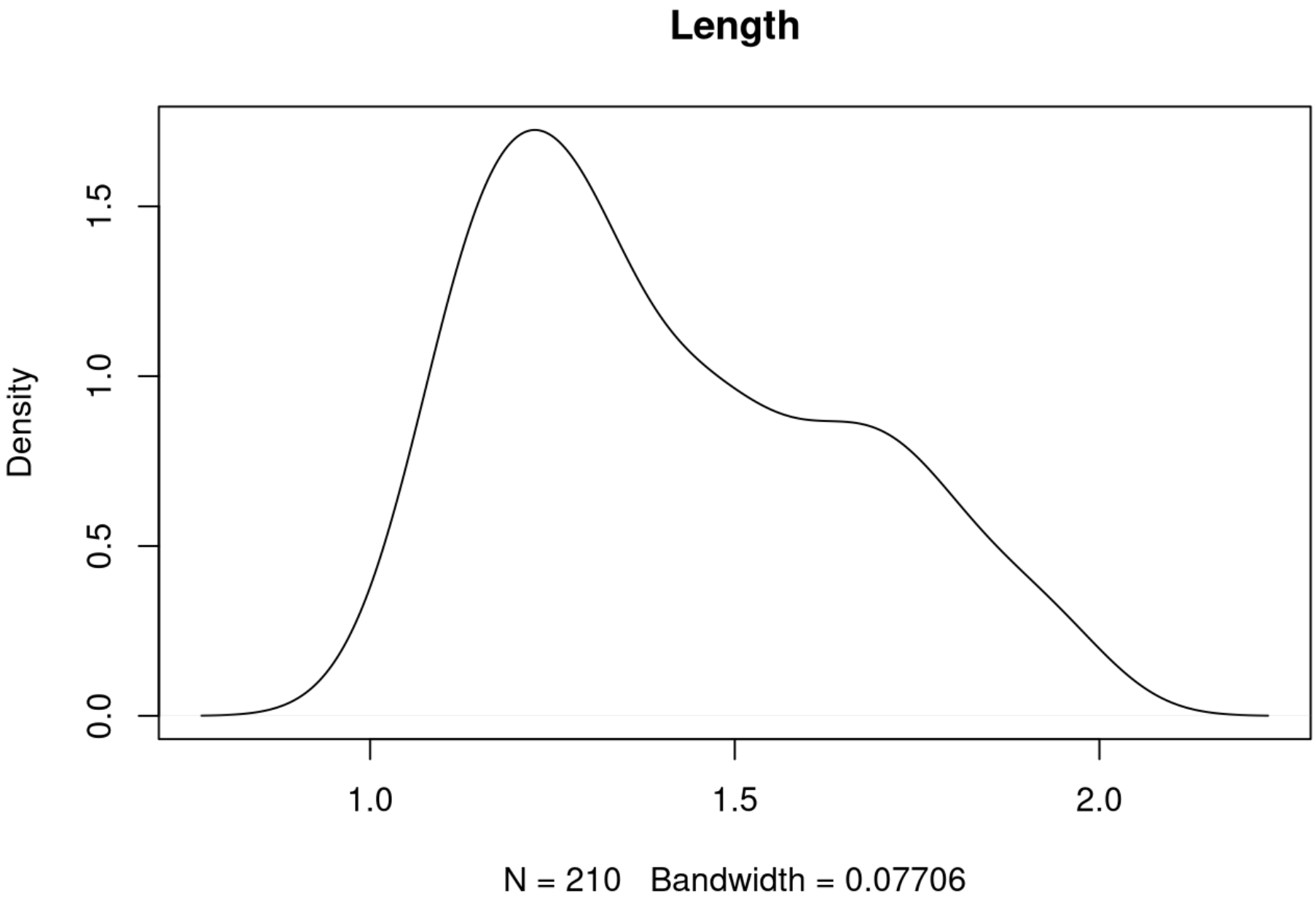

)


Compactness
Density
0.0
0.5
1.0
1.5
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
N = 210 Bandwidth = 0.06458

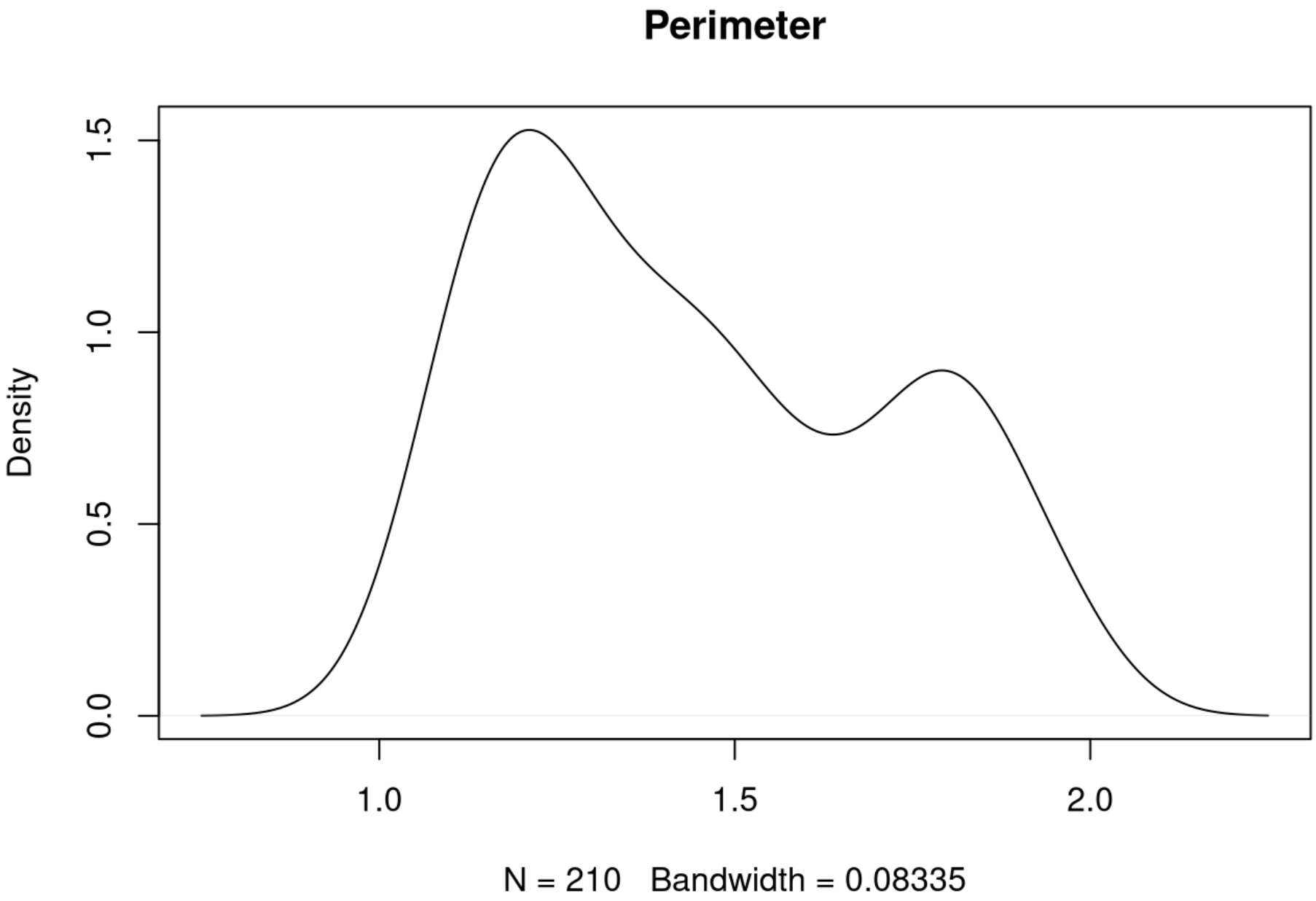
)


Groove.Length
Density
0.0
0.5
1.0
1.5
2.0
1.0
1.5
2.0
N = 210 Bandwidth = 0.07475

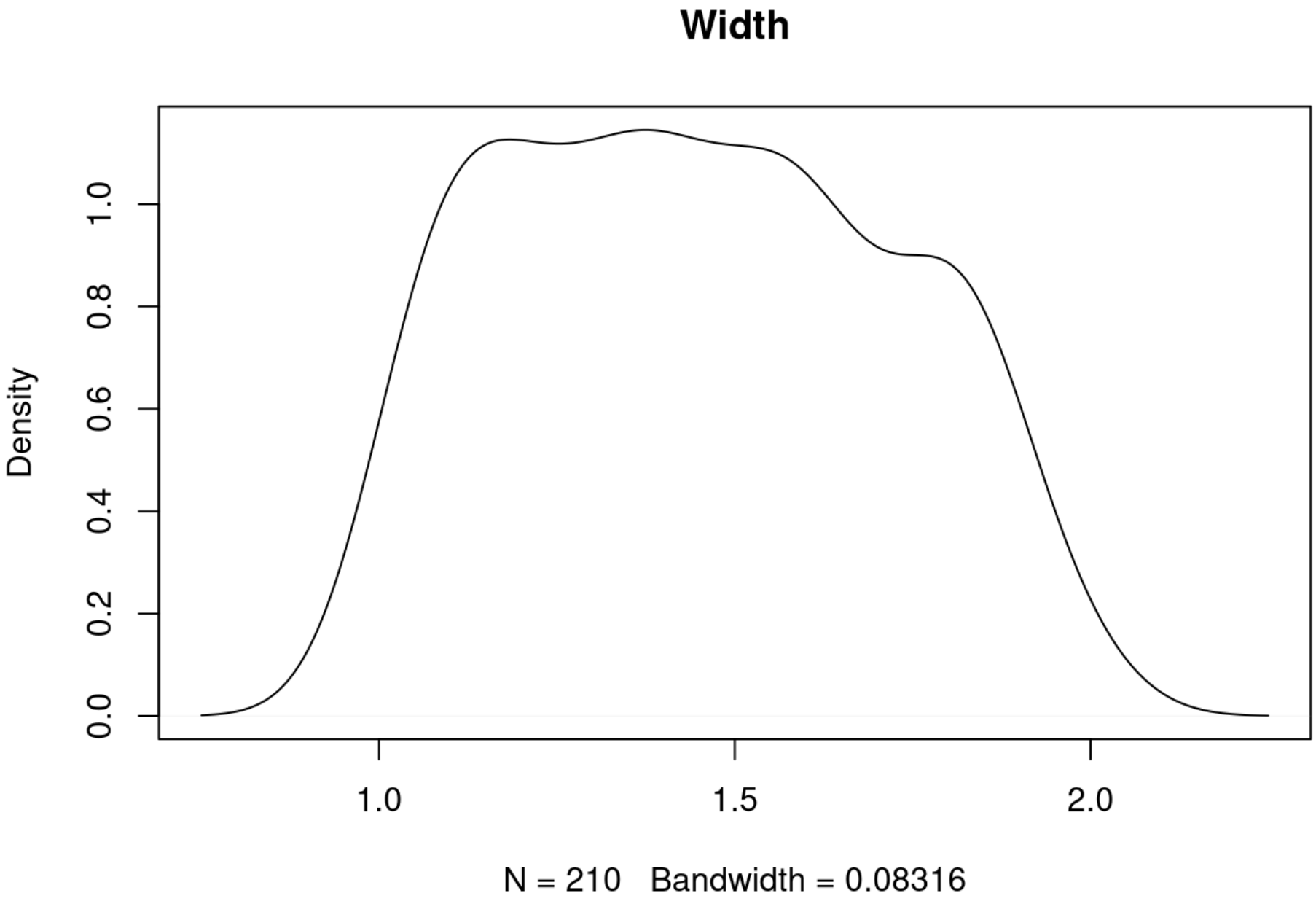
)


Length
Density
0.0
0.5
1.0
1.5
1.0
1.5
2.0
N = 210 Bandwidth = 0.07706

)


Perimeter
Density
0.0
0.5
1.0
1.5
1.0
1.5
2.0
N = 210 Bandwidth = 0.08335

)


Width
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
N = 210 Bandwidth = 0.08316

)

**Experiment Table Example**

| Sample Size | Layer Units | Model Loss | Optimizer | Accuracy | Epochs | Input Shape | Feature Selection | Test Loss | Test Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| 150 | 128 | categorical_crossentropy | adam | accuracy | 75 | 3 | Area, Asymmetry, Compactness | 0.4799 | 0.8 |