Multivariate Analysis

# Shape Features

| For each  Nucleus (A),  Cytoplasm (B) | **'area', 'extent', 'solidity', 'RAN', 'convexity', perimeter, 'major\_axis', minor\_axis, equivalent\_diameter, 'radial\_distance\_maxN', 'radial\_distance\_mean, 'radial\_distance\_sd', eccentricity, circularity, RI, 'radial\_distance\_Entropia, 'radial\_distance\_kurtose',**  **Use\_curv1,Use\_curv2, 'Use\_curv3, major\_axis\_angle** |
| --- | --- |
| For each  Cell (C) | **area\_NC', 'perimetro\_NC', 'major\_axis\_NC', 'minor\_axis\_NC', 'nucleus\_position, sub\_major\_axis\_angle\_NC',convexity\_NC'** |

The following procedure was executed in Python to reduce the amount of variables, as follow:

* Correlation matrix analysis to exclude features with high degree of correlated values in each group: Nucleus, Cytoplasm, Cell
* Analysis of boxplots for the remaining features and their values distributed per Bethesda classes. All features with similar distribution among normal and grade classes, confirmed by a T-Test, was excluded.
* Analysis of a pairwise distribution of all remaining features from nuclei, cytoplasm and cells groups together.
* Classification using a KNN

## Núcleus features

### Correlation matrix analysis

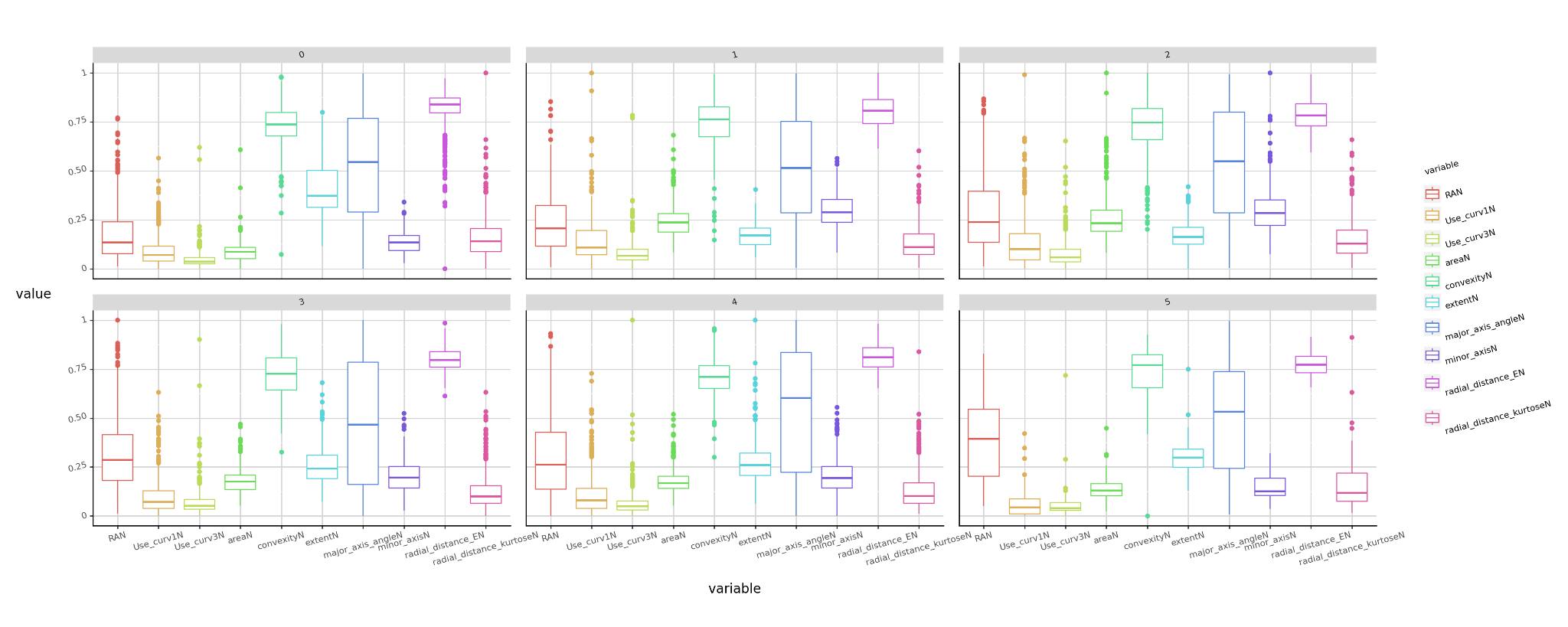
## 

**Consideration:** for each group of features highly correlated (> 90% ) the highlighted feature was selected, all others were excluded.

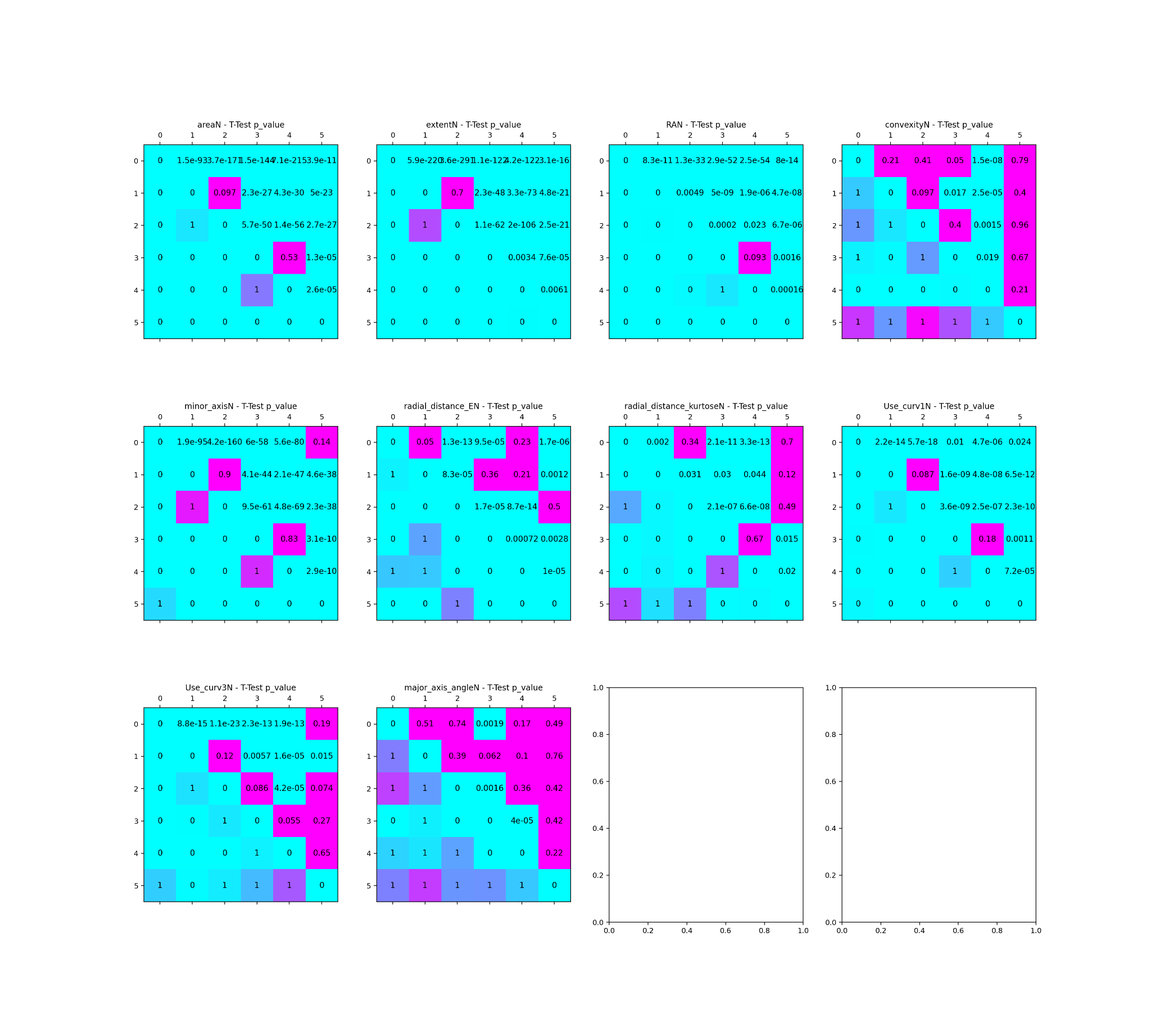
* **area**, perimeter, major\_axis, equivalent\_diameter, radial\_distance\_max ,radial\_distance\_mean.
* **extent** (Ratio of pixels in the region to pixels in the total bounding box), solidity (Ratio of pixels in the region to pixels of the convex hull image), circularity (perimeter^2/(4pi\*Area)), RI (rugosity index).
* **RAN** (ration between: sum of radial distances bigger than mean / (N \* mean radial distance), radial distance standard dev, eccentricity (convex Hull perimeter /perimeter)

### Boxplots distribution per Bethesda classes

### 



### T-Test



**Consideration:** the follow attribute will be excluded to avoid confusions among normal and abnormal cells or among low grade lesions and high grade lesions

* 1. Convexity: 0(Normal), 1 (ASCUS), 2(LSIL), 3(ASCH), 5(SCC)
  2. Minor\_axis: 0(Normal), 5(SCC)
  3. radial\_distance\_Entropy: 2(LSIL) e 5 (SCC)
  4. radial\_distance\_kurtose: 1(ASCUS) e 4(HSIL)
  5. Use\_curv3
  6. major\_axis\_angle

### **Nucleus** features selected:

#### area,

#### extent Ratio of pixels in the region to pixels in the total bounding box

#### RA Sum of radial distances bigger than mean distance / (N \* mean distance)

1. Use\_curv1 the circle area having the same radio of the curvature divided by the area

### 

## Cytoplasm features

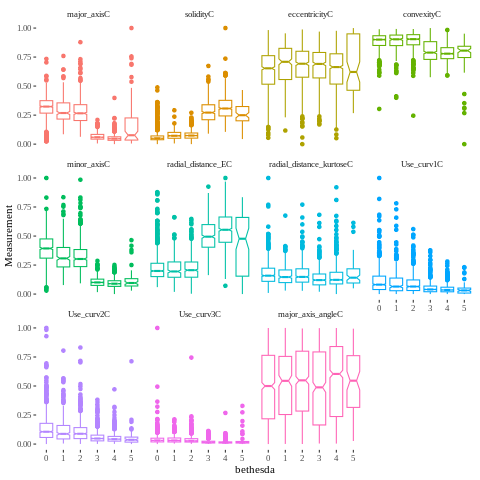
### Correlation matrix analysis

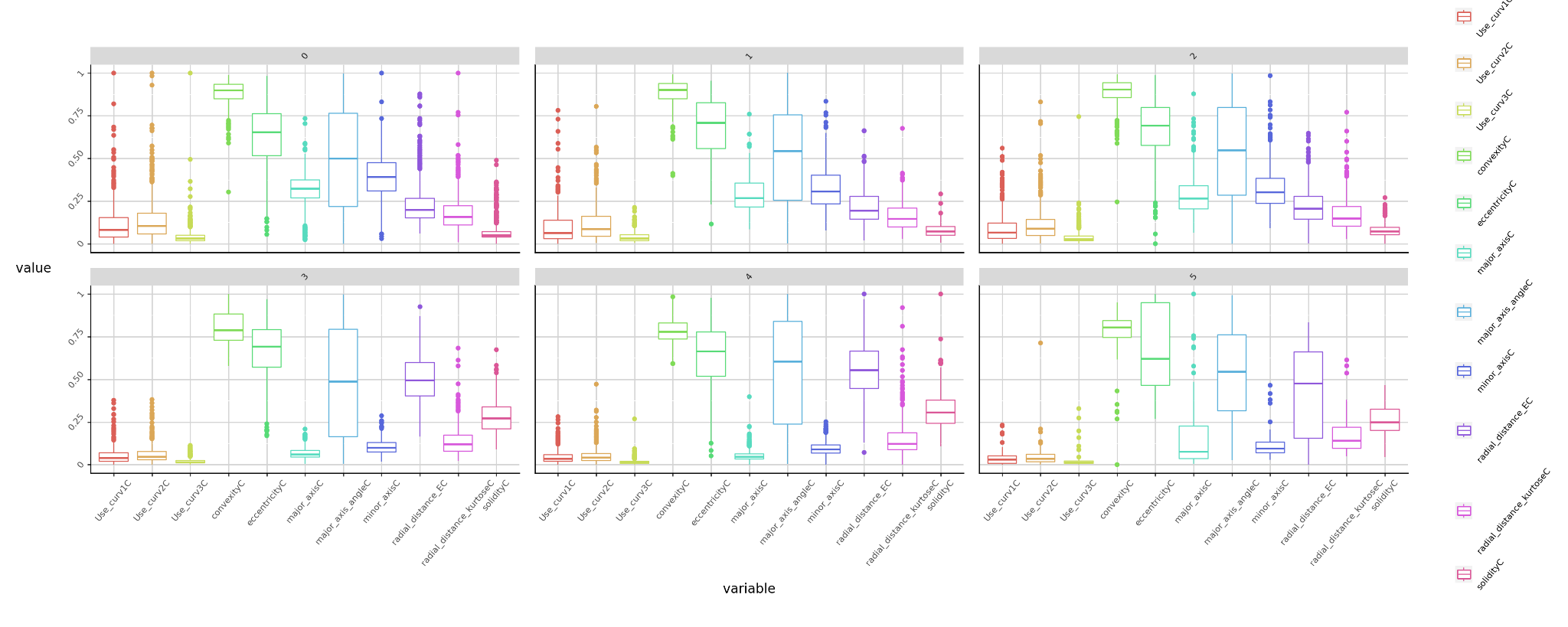


**Consideration:** here we observed the same pattern of nucleus matrix. Again, for each group of features highly correlated (> 82% ), the highlighted feature was selected, all others were excluded.

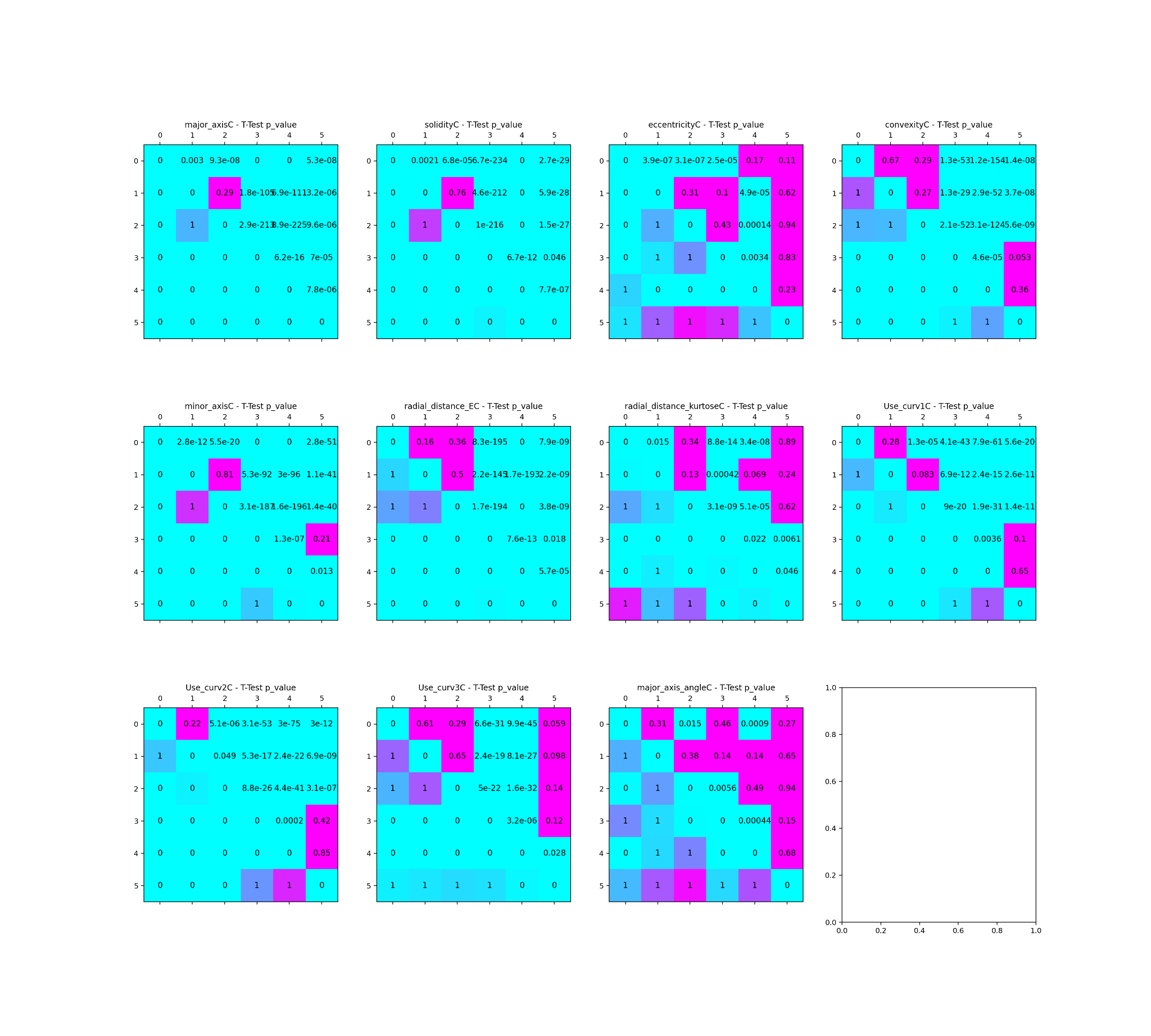
1. **major\_axis,** area, perimeter, equivalent\_diameter, radial\_distance\_max, radial\_distance\_mean.
2. **solidity**, extent, (Ratio of pixels in the region to pixels of the convex hull image), circularidade (p^2/($pi\*A)), RI (rugosity index).
3. **eccentricity** (perimetro\_convexHull/perimetro), RA, radial\_distance\_sd,
4. **Use\_curv1**, Use\_curv2.

### Boxplots distribution per Bethesda classes





### T-Test



**Consideration:** the follow attribute will be excluded to avoid confusions among normal and abnormal cells or among low grade lesions and high grade lesions

1. eccentricity
2. convexity
3. minor\_axis
4. radial\_distance\_entropy
5. radial\_distance\_kurtose
6. Use\_curv1, Use\_curv2, Use\_curv3
7. major\_axis\_angle

The remaining features presents similar distributions among NILM and Low grade classes (e.g. 'major\_axis', 'solidity , 'minor\_axis)

### **Cytoplasm** features selected still presents similar distributions among normal and low grade classes

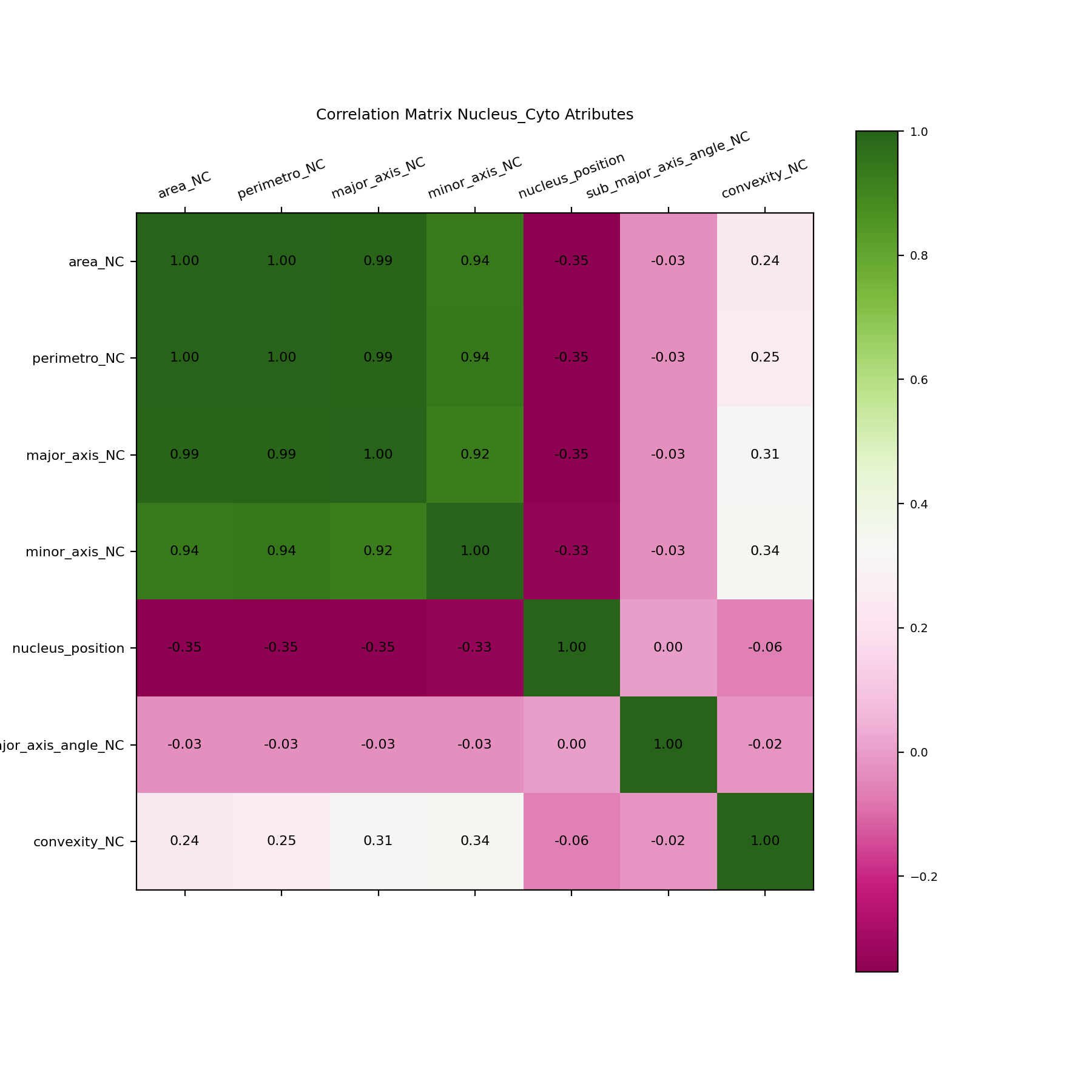
#### solidity (Ratio of pixels in the region to pixels of the convex hull image)

#### major axis

#### 

## Cells features

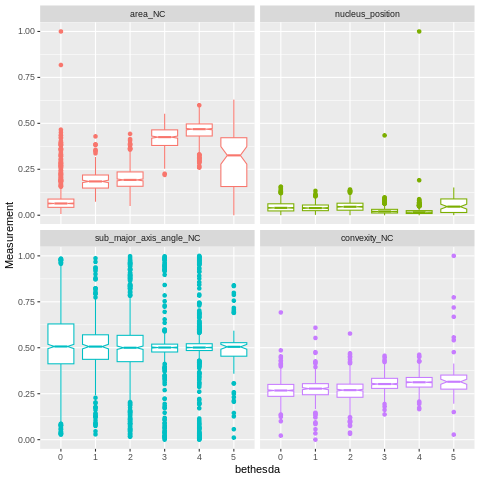
### Correlation matrix analysis

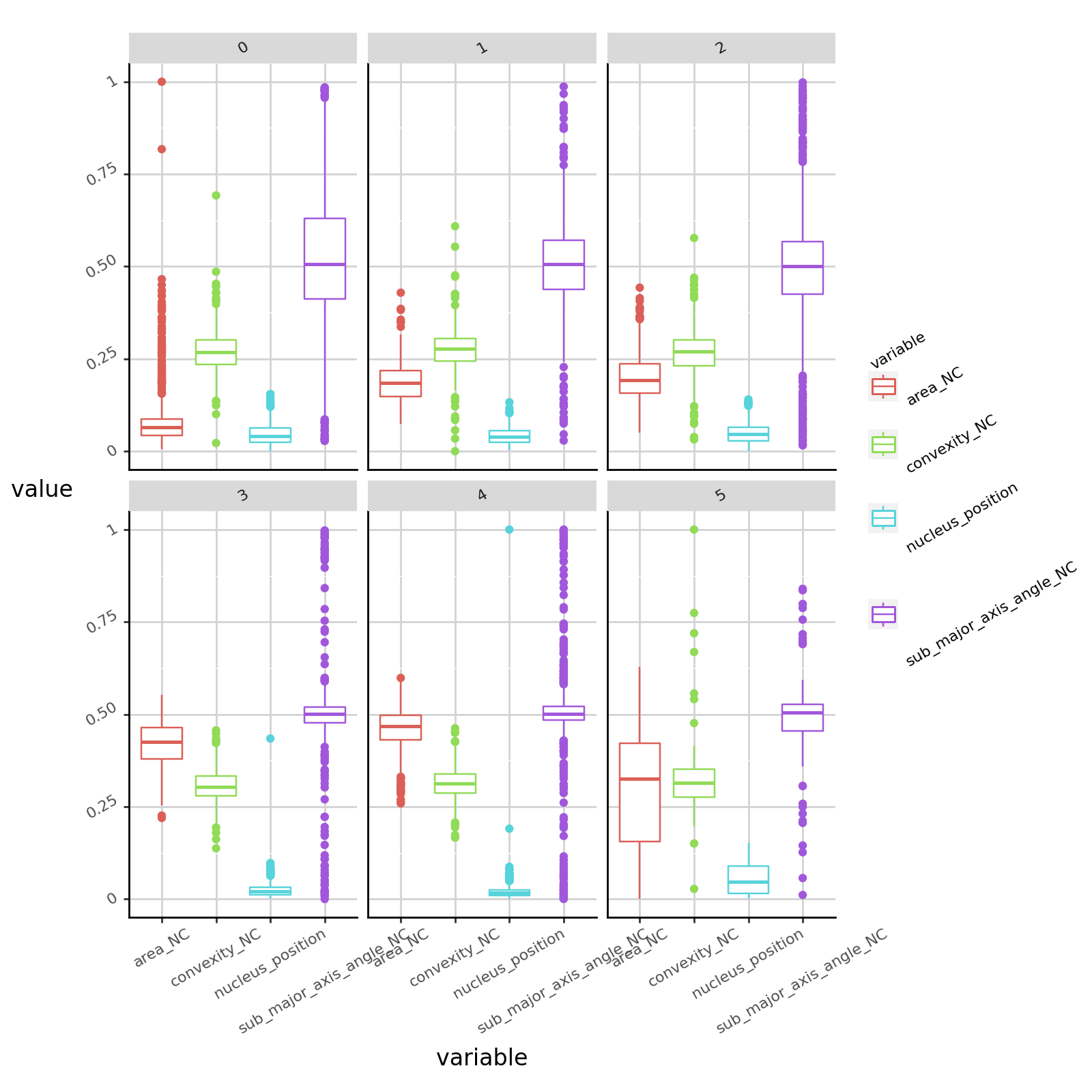


**Consideration:** the above matrix shows high correlation among the features representing a size ratio between the nucleus and cytoplasm :

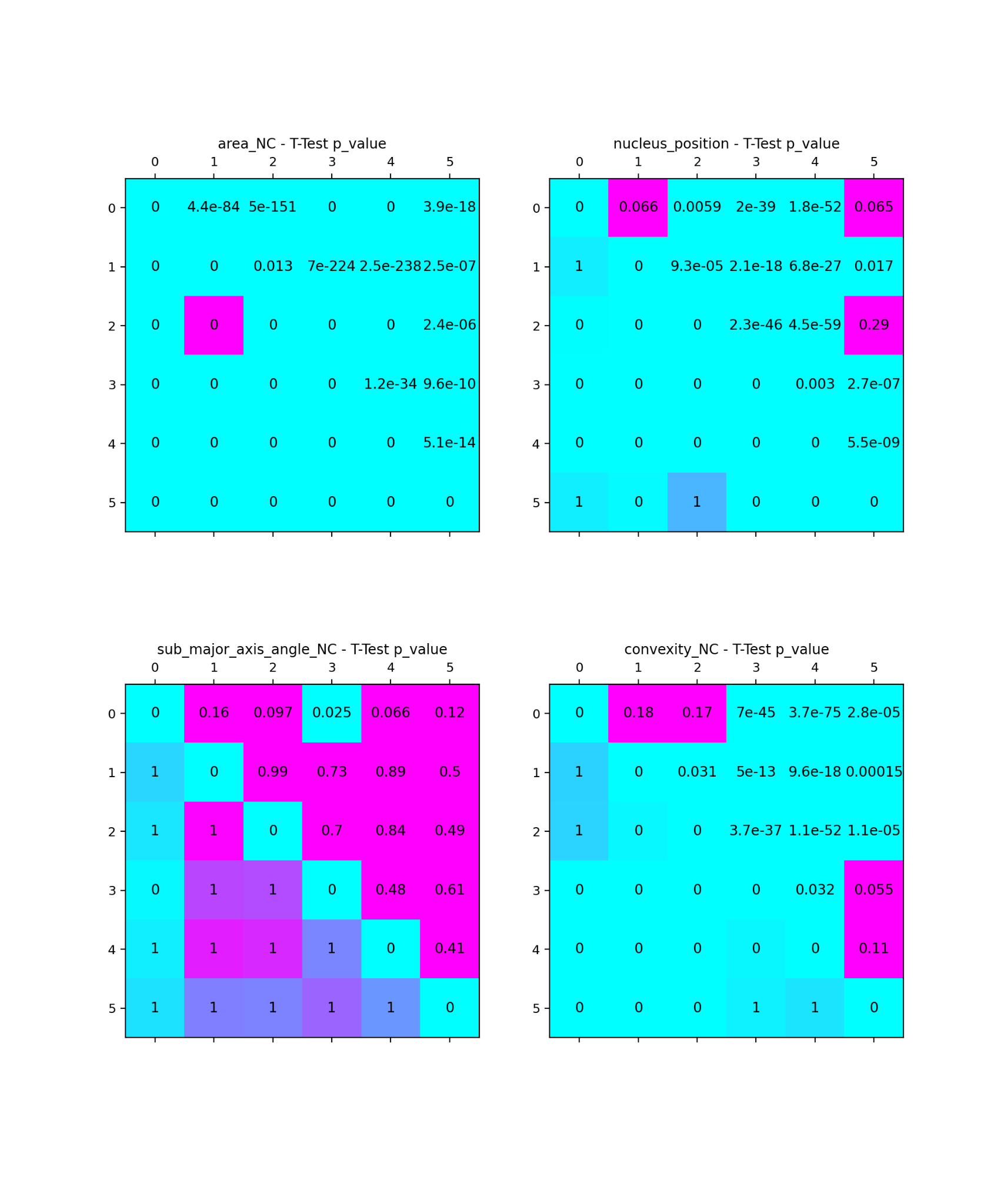
**area\_NC**  (will be used) , perimeter\_NC, major\_axis\_NC , minor\_axis\_NC

### Boxplots distribution per Bethesda classes





### T-Test



**Consideration:** the follow attribute will be excluded to avoid confusions among normal and abnormal cells or among low grade lesions and high grade lesions

* nucleus position
* sub\_major\_axis\_angle
* convexidade

### **Cell** feature selected

#### **Area\_NC** (ration of nucleus and cytoplasm areas)

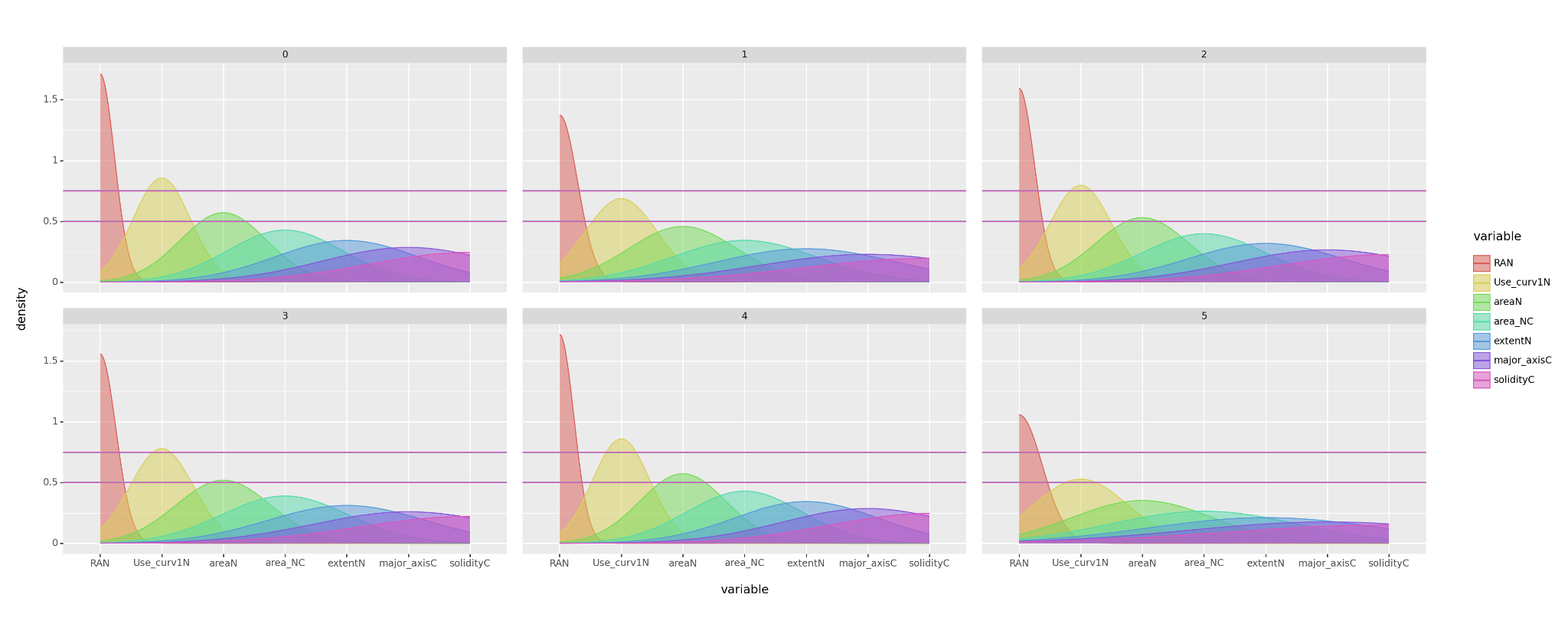
# Analysis of remaining nuclei, cytoplasm and cells

### Remaining Features:

| Nucleus (A) | **area, extent (Ratio of pixels in the region to pixels in the total bounding box), RA, Use\_curv1** |
| --- | --- |
| Cytoplasm (B) | **solidity (Ratio of pixels in the region to pixels of the convex hull image), major\_axis** |
| Cell (C) | **area\_NC** |

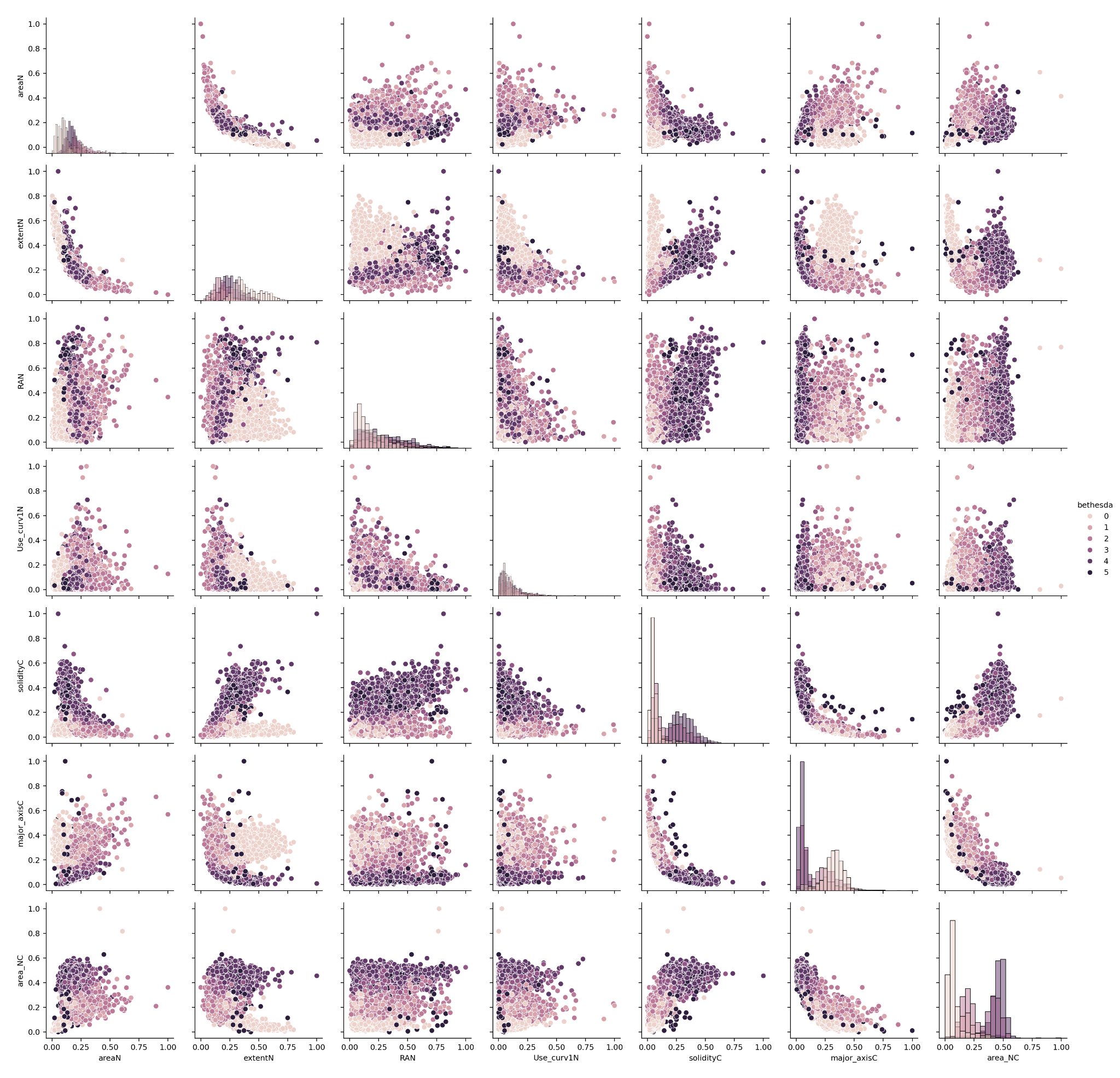
### 

### Density distribution



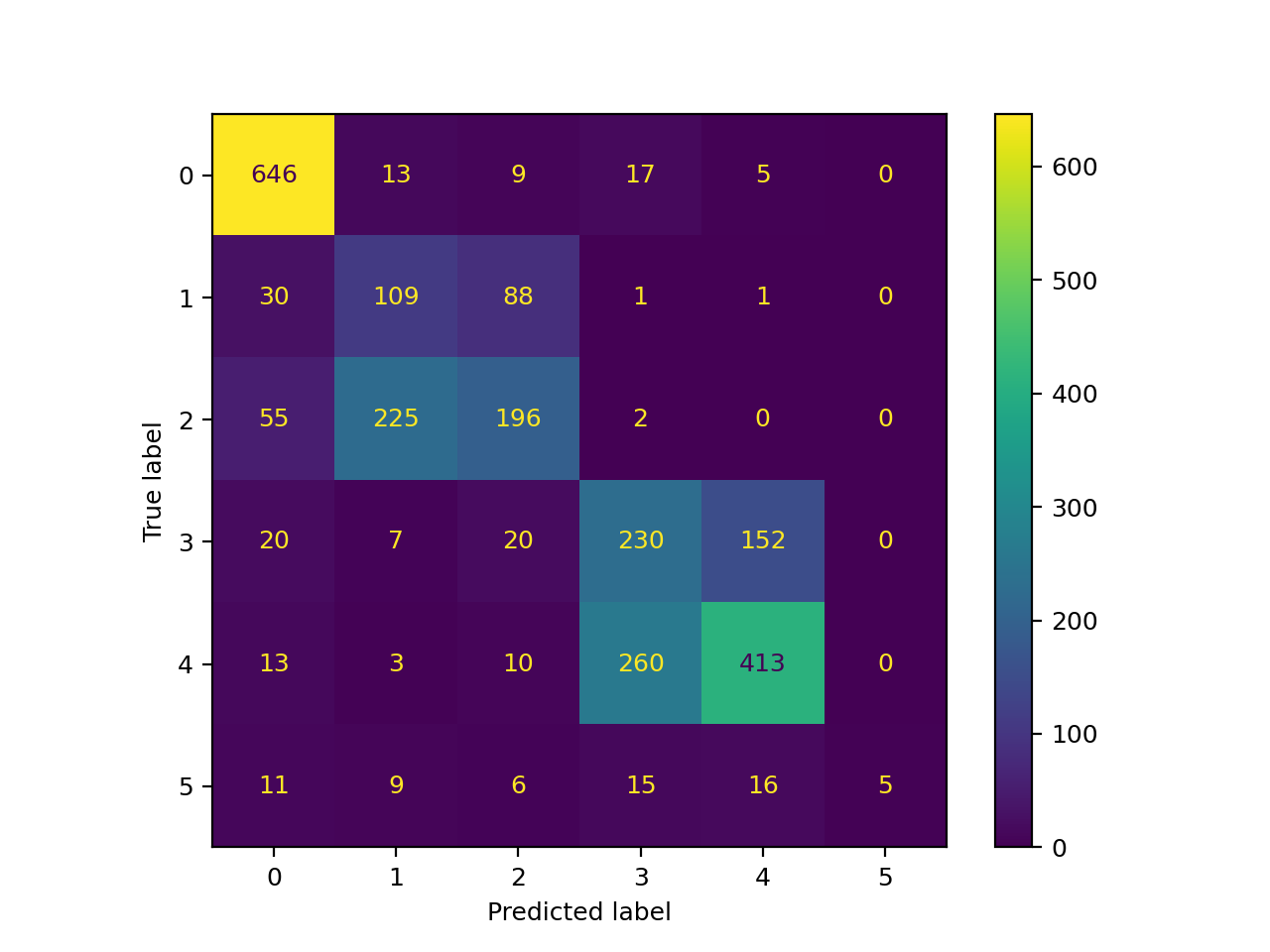
**Consideration:** very similar distributions  **0-4 (normal and HSIL), 2-3 (ASCUS and ASCH)**

### Pairwise density distribution



# Classification (KNN, SVM)

* KNN (2 neighboors): 0.6180904522613065 with 80% of test data



* SVN: 0.7125193199381762 with 40% of test data

