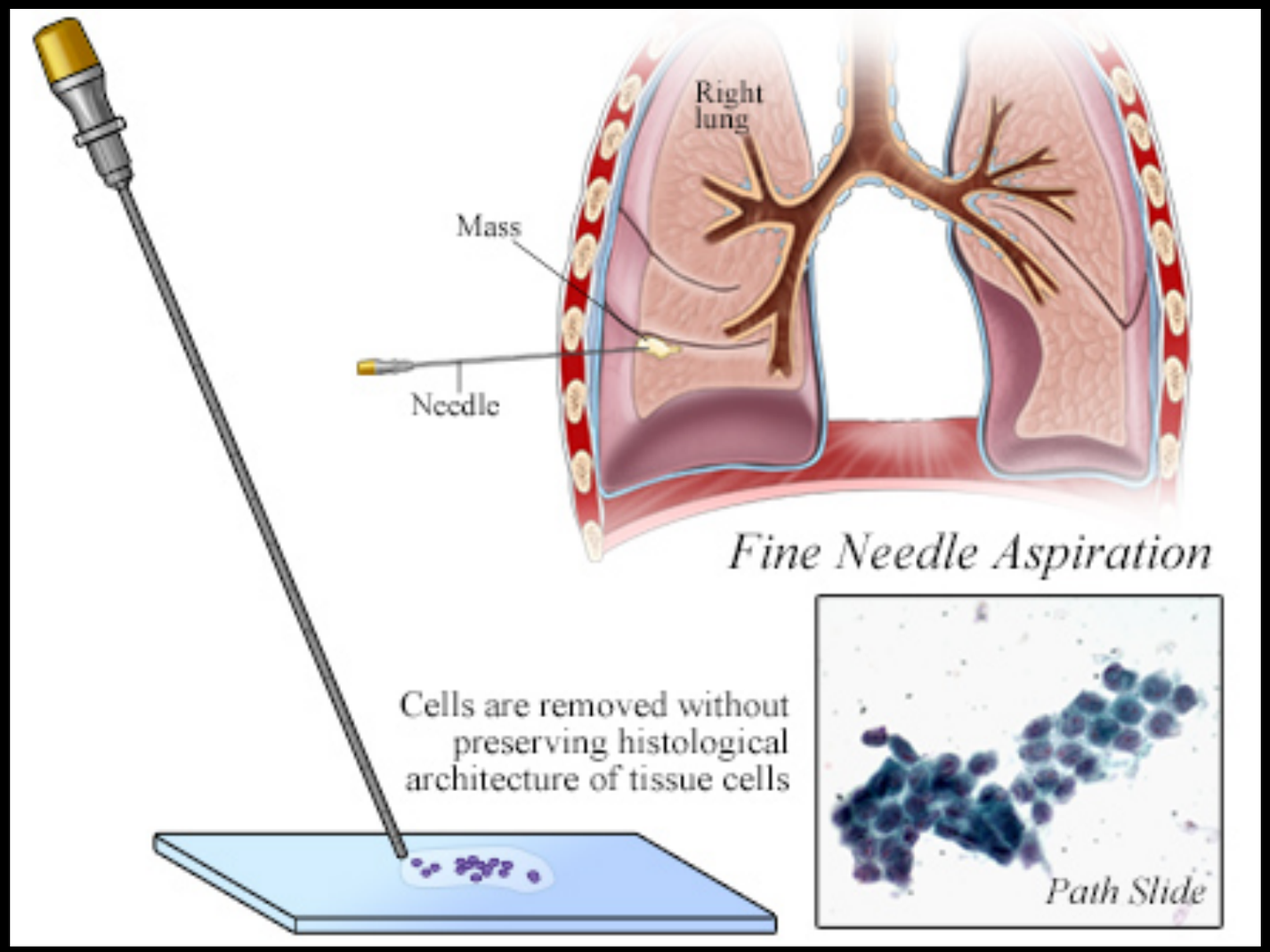
Data Storytelling of Univ of Wisconsin Breast Cancer Data Set

Murali Satuluri

Problem



- **Fine Needle Aspiration is a less invasive alternative to Biopsy.**
- **Cells collected from this test are studied and their features are recorded.**
- **The features of the cell are to be used to predict if the parent tissue is malignant or benign.**

Variables

Input Variables

1.Clump Thickness

2.Uniformity of Cell Sizes

3.Uniformity of Cell Shape

4.Marginal Adhesion

5.Single Epithelial Cell Size

6.Bare Nuclei

7.Bland Chromatin

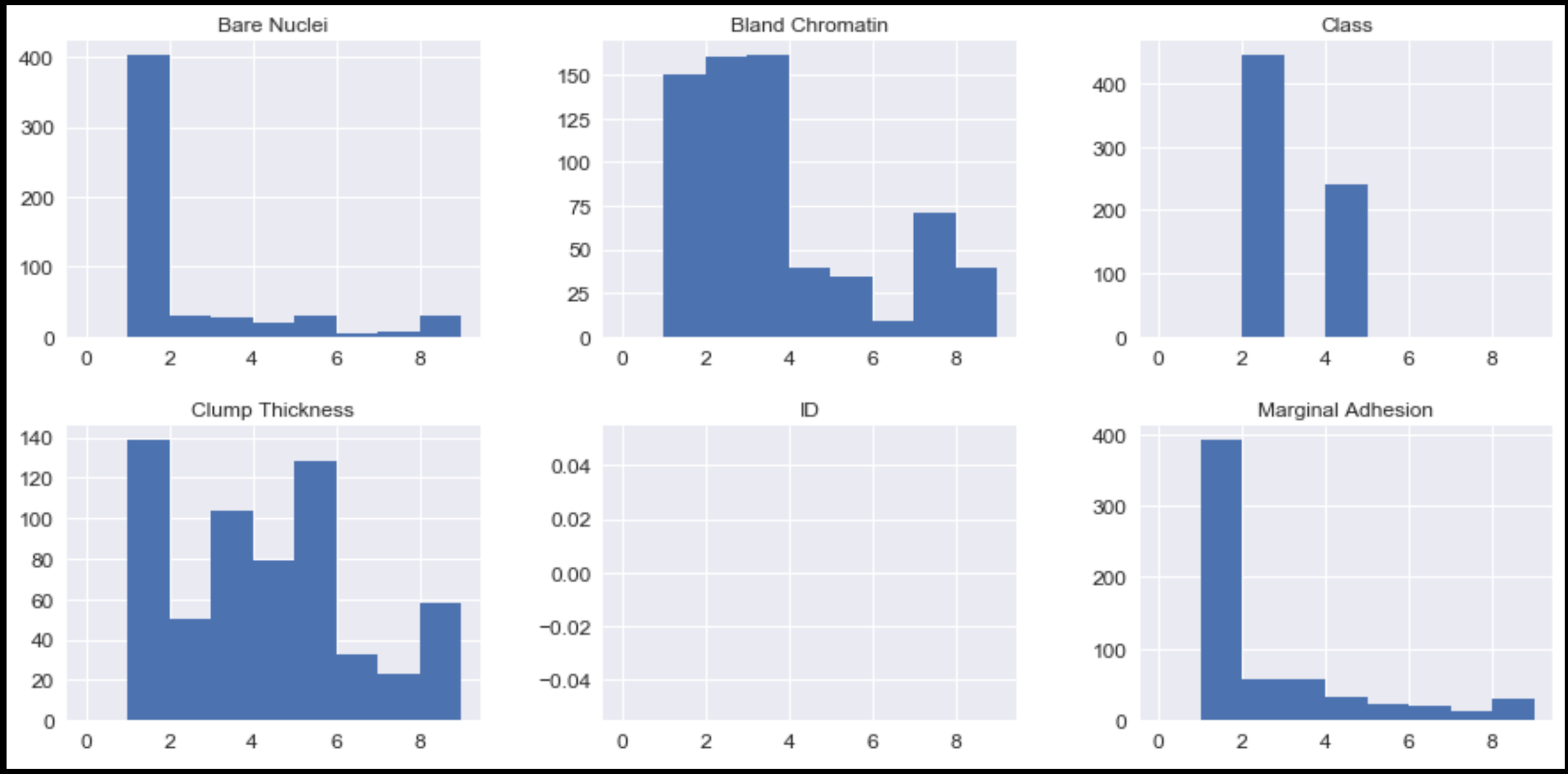
8.Normal Nucleoli

9.Mitoses

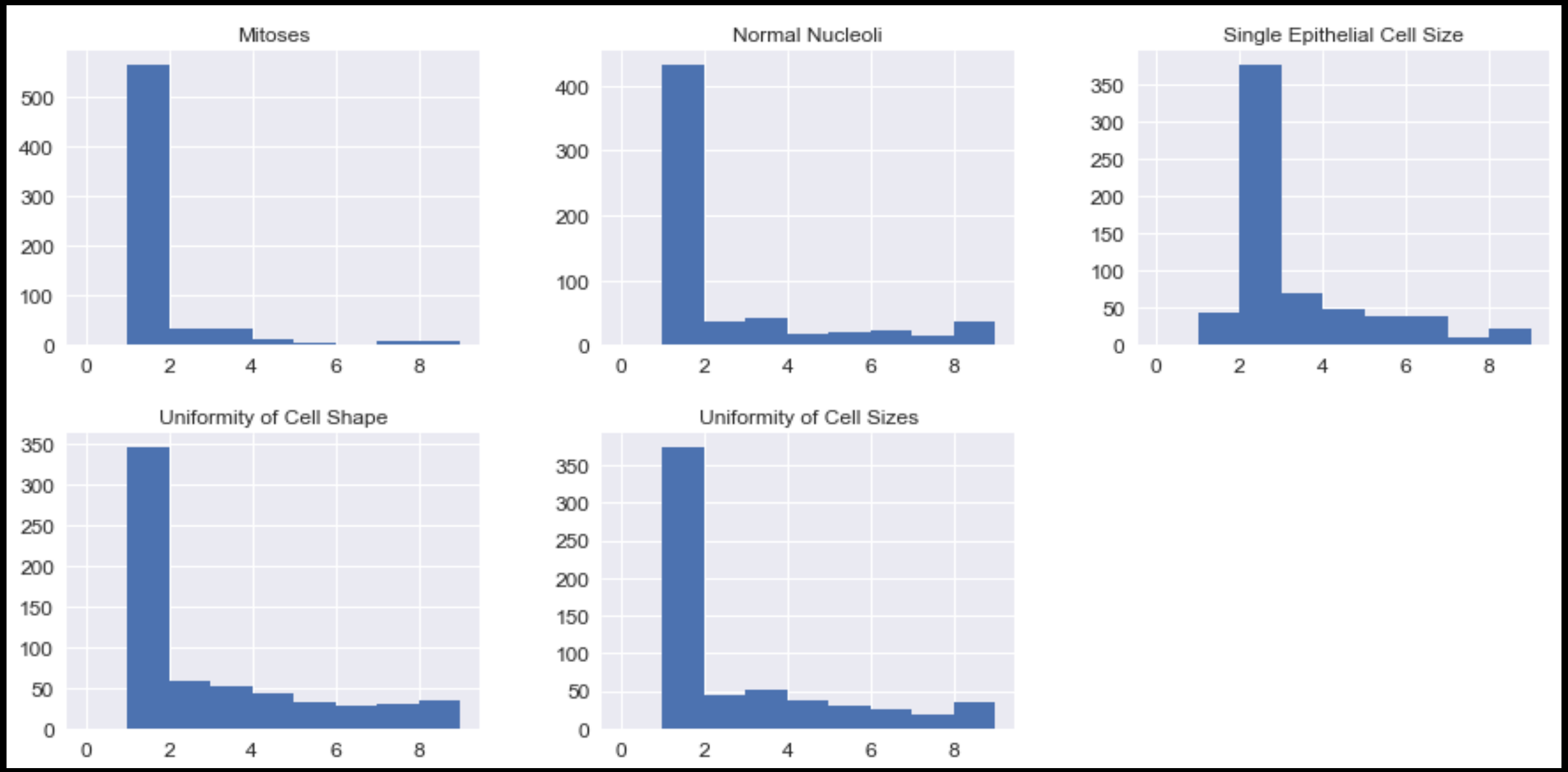
Output Variable

Cell Classification
(i.e. Benign or Malignant)

Input Variables

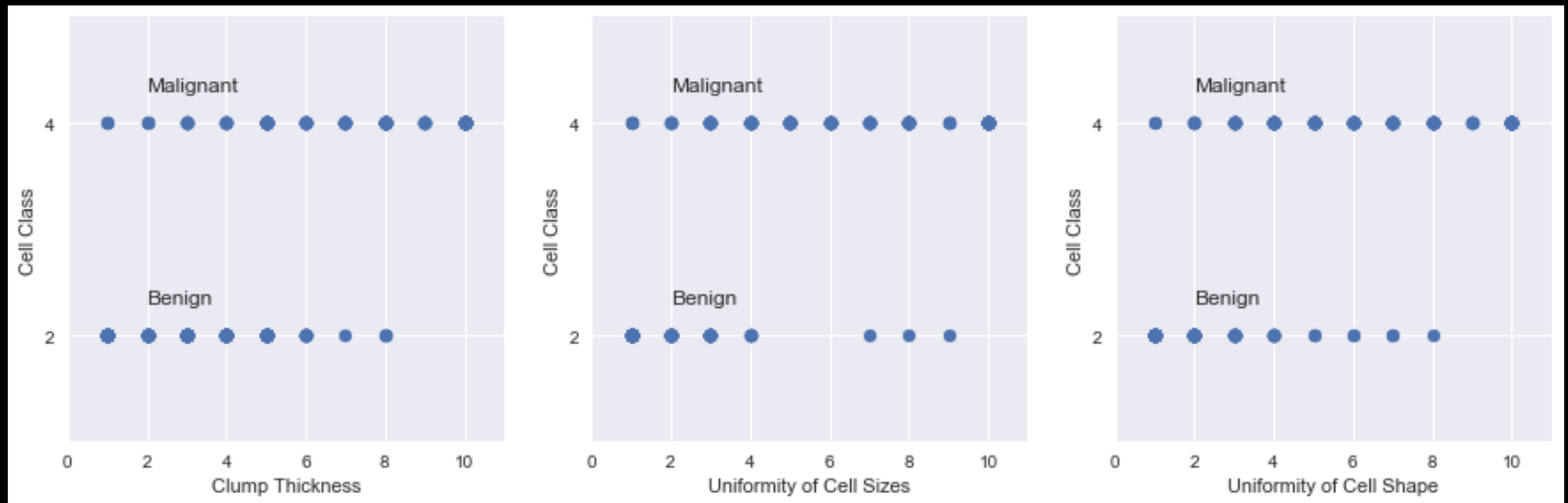


Input Variables



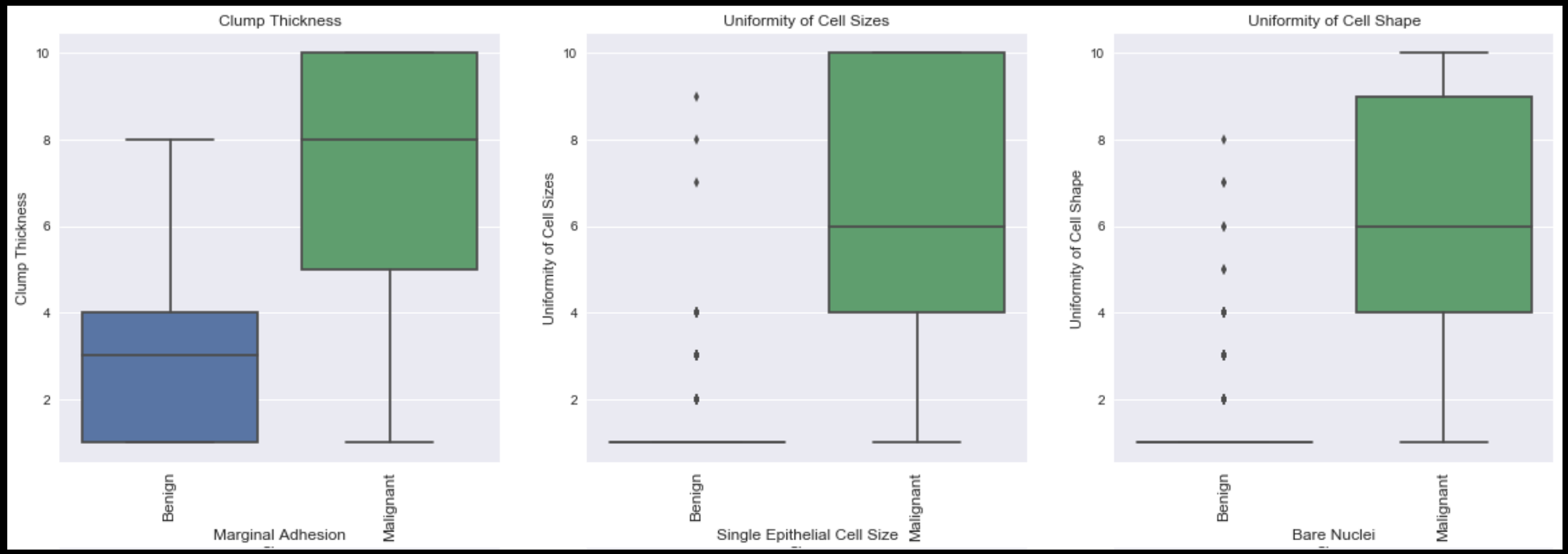
Variables are not normally distributed. They are skewed to the left.
These variables clearly seem correlated.

Scatter Plots



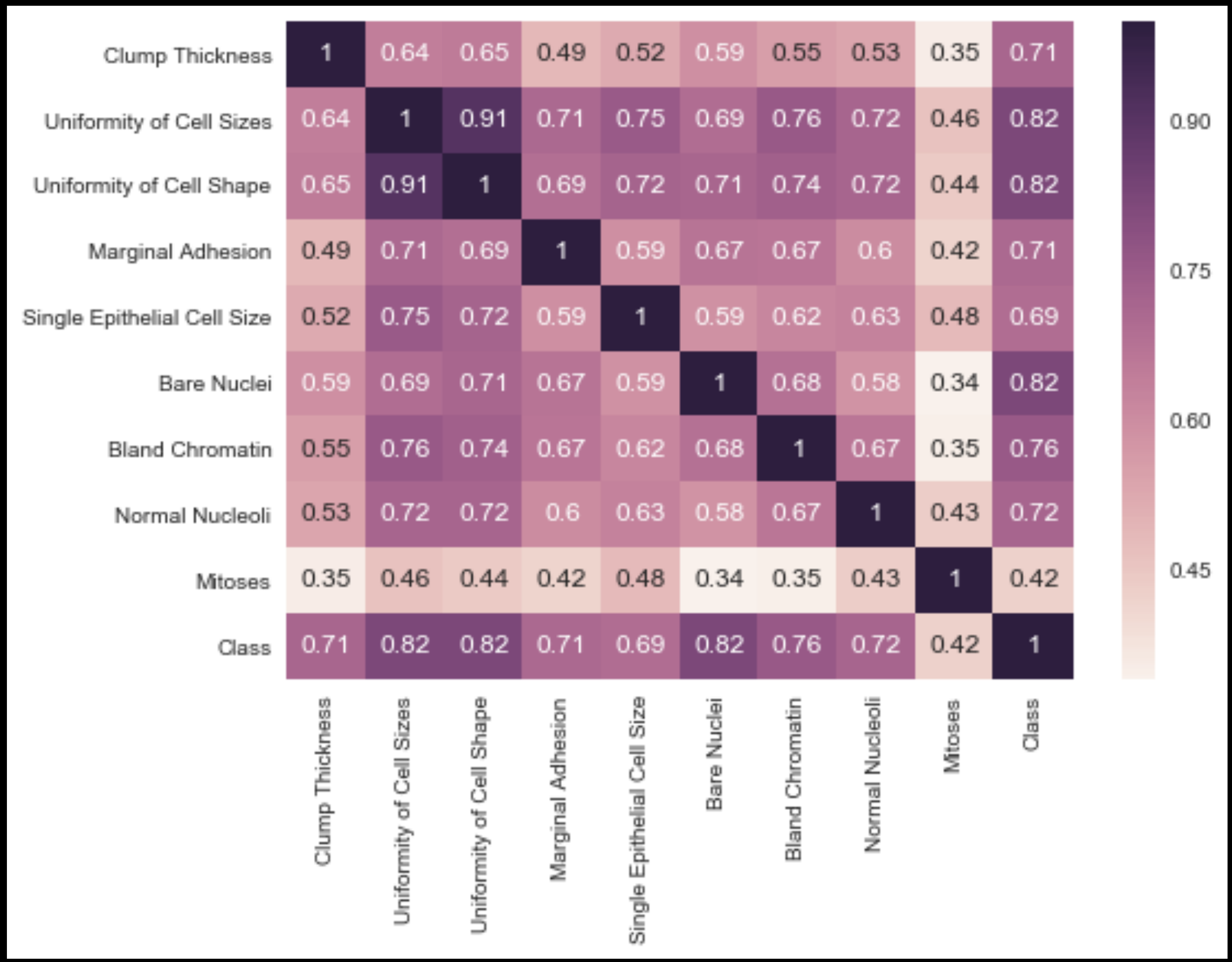
Scatter Plots don’t help

Box Plots



Box plots filtered by Cell Type Classification help see the impact of various input variables on the output variable.

Correlation Matrix



Conclusions

- **Most of the input variables are correlated to the ‘Uniformity of the cell size’.**
- **Scatter Plots are not helpful for evaluating relations between qualitative variables**
- **Cell’s tend to be benign if the ‘Uniformity of cell size’ is lower in value.**