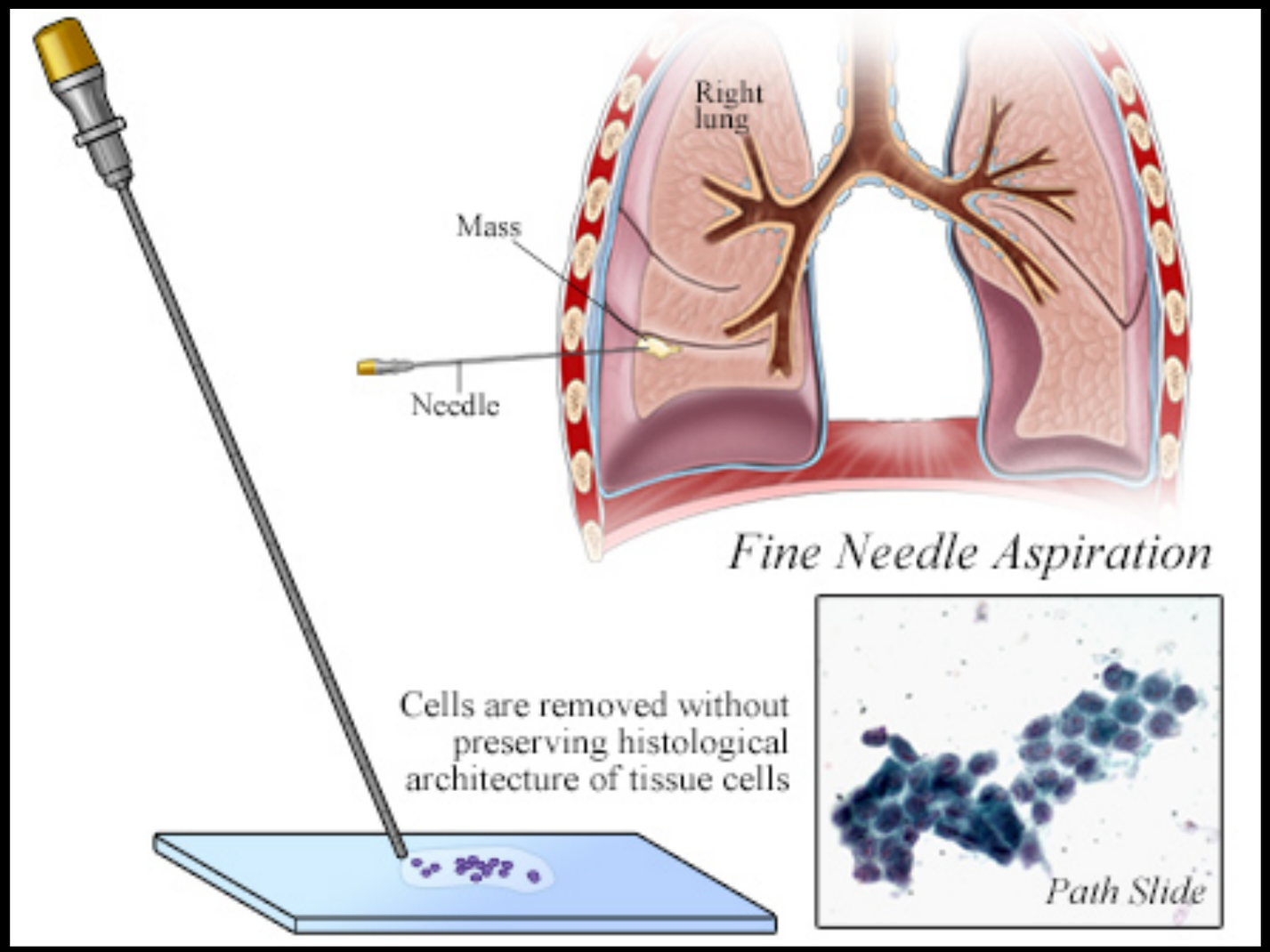
Data Storytelling of Univ of Wisconsin Breast Cancer Data Set

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

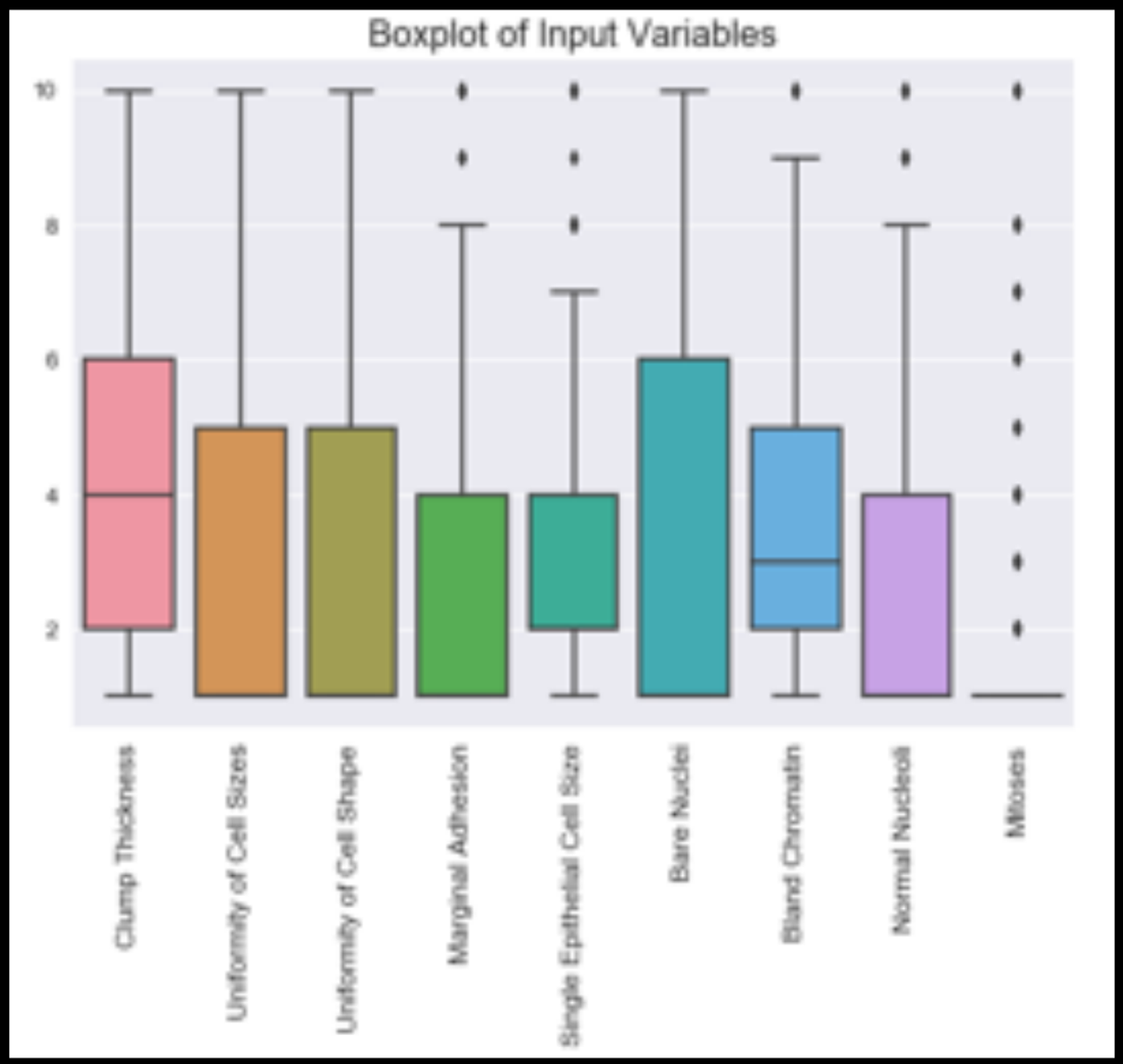
9.Mitoses

Output Variable

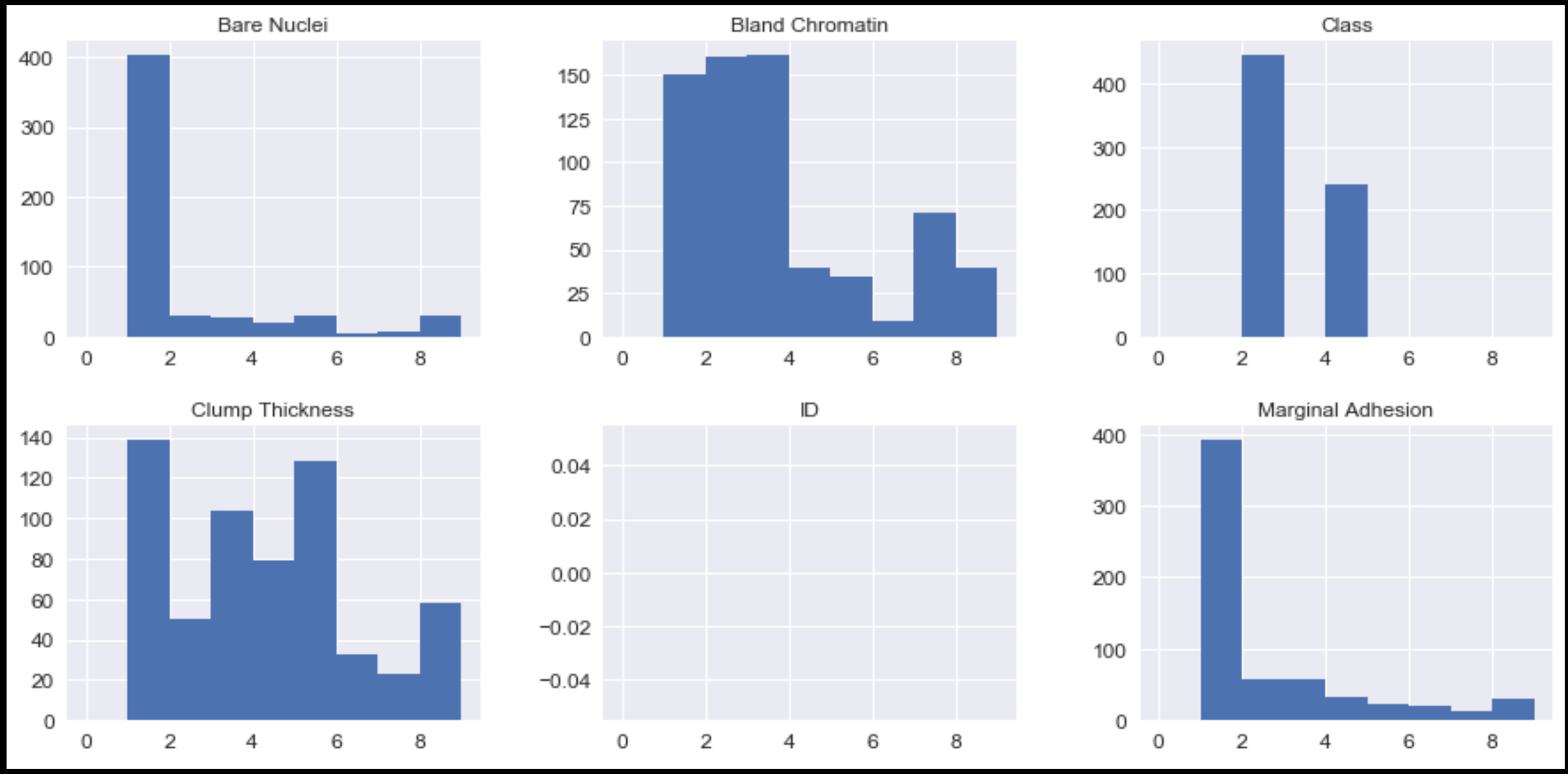
Cell Classification
(i.e. Benign or Malignant)

Data Cleaning

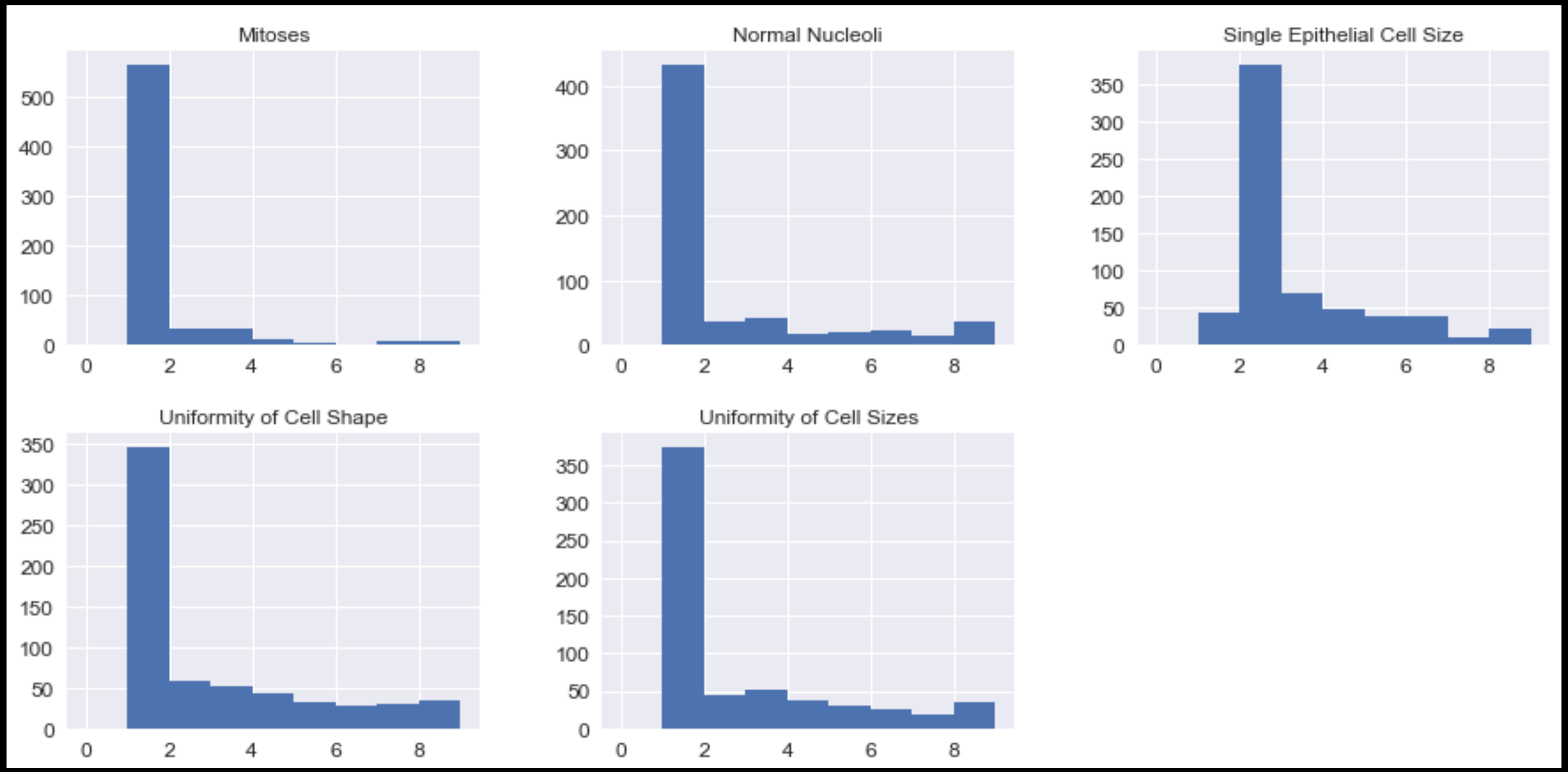
- There are 699 samples. 16 of them containing NA values were removed.
- Boxplots of the remaining data are shown below.
- Variables are not normally distributed. Since there is no clear basis on which the extreme data can be called outliers, no further data were removed.



Input Variables

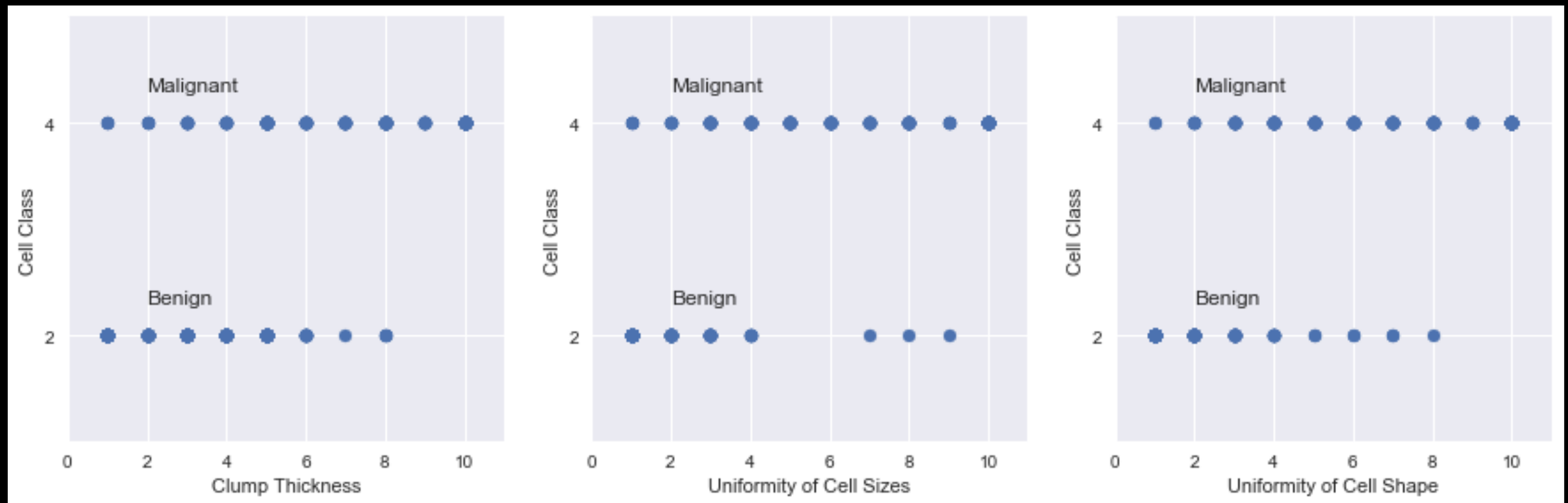


Input Variables



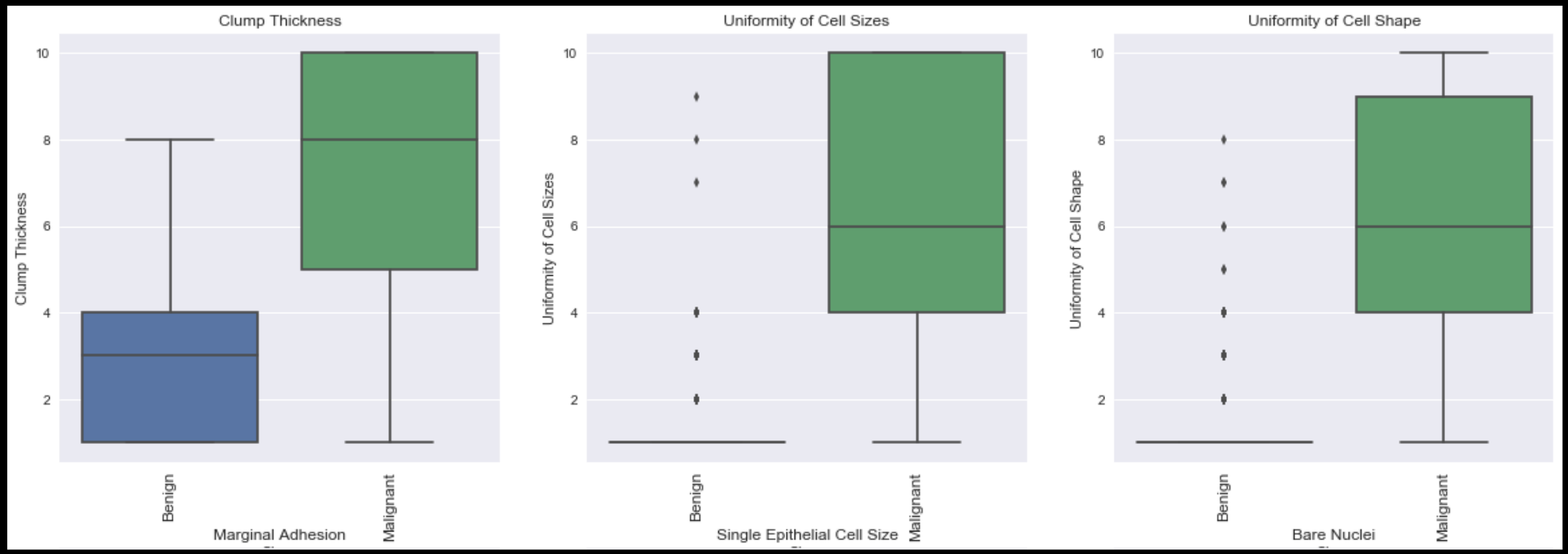
Variables are not normally distributed. They are skewed to the left.
These variables clearly seem correlated.

Scatter Plots



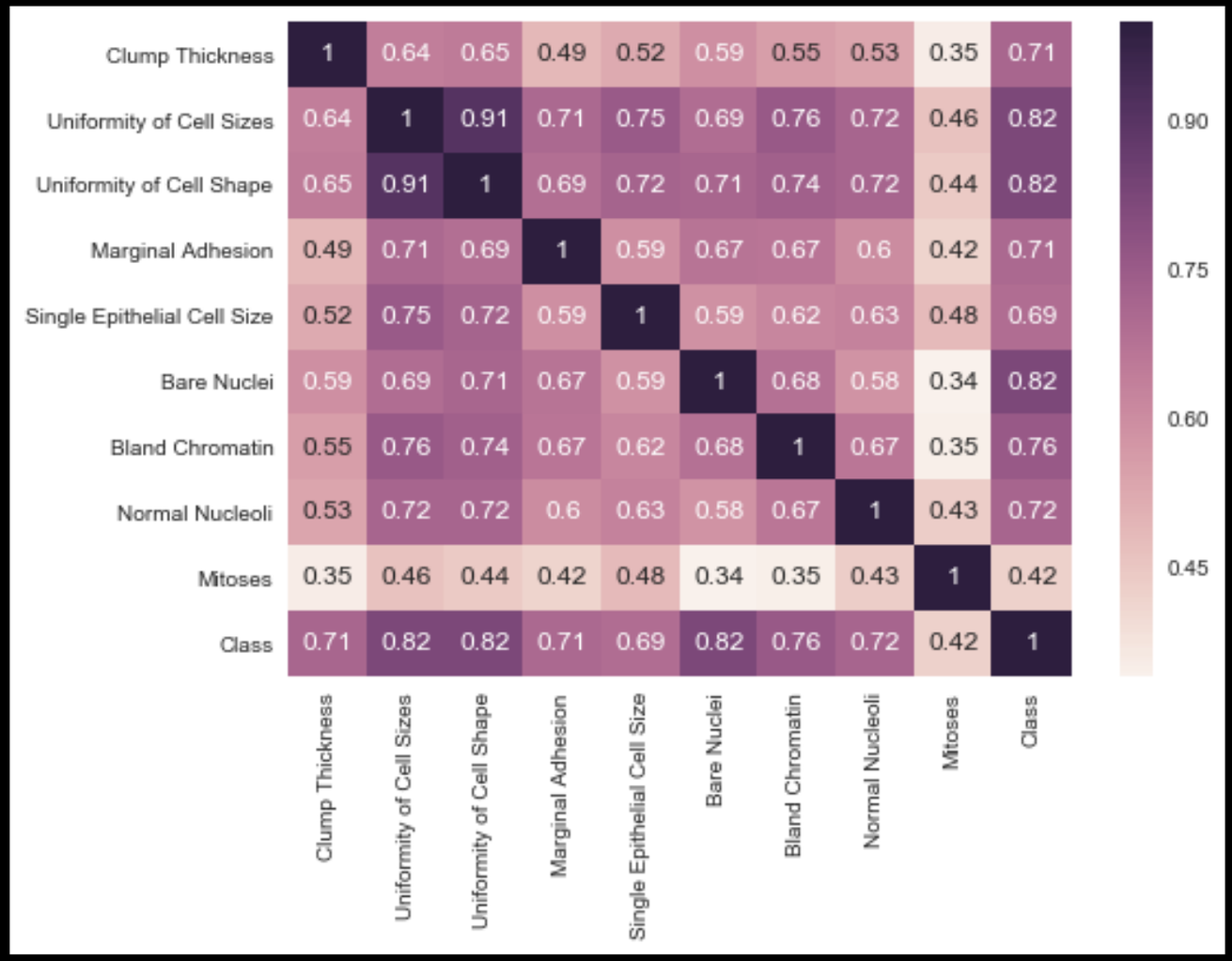
Scatter Plots don’t help

Box Plots



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

Correlation Matrix



- **Most of the input variables are correlated to the ‘Uniformity of the cell size & shape’.**

- **There is also good correlation between the cell class & input variables.**

Logistic Regression Model

Linear Logistic Regression Model was built using all variables.
This model has 97% Accuracy.

Cell Type	Actual Benign	Actual Malignant
Predicted Benign	47	0
Predicted Malignant	1	21

Variable Name	Coefficients	Variable Name	Coefficients
Bare Nuclei	0.30	**Marginal Adhesion**	0.19
Uniformity of Cell Shape	0.27	**Bland Chromatin**	0.19
Clump Thickness	0.25	**Uniformity of Cell Sizes**	0.16
Mitoses	0.22	**Normal Nucleoli**	0.15
		Single Epithelial Cell Size	-0.04

- **Cell's tend to be benign if the input variables are lower in value.**
- **Build a model using the most significant variables (underlined in the above table).**

Logistic Regression Model

Several bootstrap runs were made on the Linear Logistic Regression Model using the 3 most important variables. Confidence intervals from these runs are shown below.

	Percentiles		
	2.5%	50%	97.5%
Bare Nuclei Coeff	0.41	0.44	0.50
Uniformity of Cell Shape Coeff	0.59	0.64	0.73
Clump Thickness Coeff	0.28	0.32	0.36
Model Accuracy %	91.30	95.65	100.00

Conclusions

- **Cell attributes from Fine Needle Aspiration procedure can be used to predict if the cell was malignant or benign.**
- **Cell attributes - Bare Nuclei, Uniformity of cell shape & clump thickness were found to be the most important.**
- **Model built using just the three variables listed above was able to predict the cell type with more than 90% accuracy.**