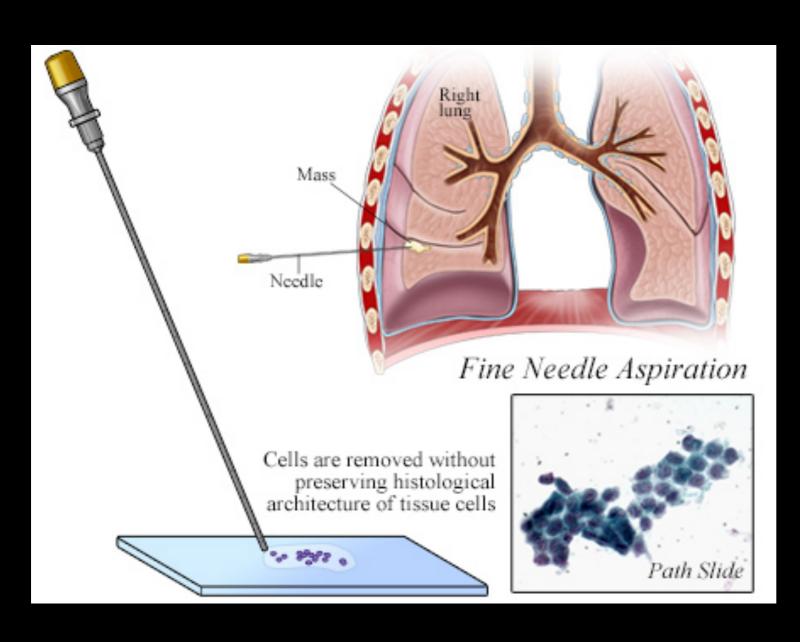
# Data Storytelling of Univ of Wisconsin Breast Cancer Data Set

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** 

**Output Variable** 

**1.Clump Thickness** 

2.Uniformity of Cell Sizes

3. Uniformity of Cell Shape

4. Marginal Adhesion

5. Single Epithelial Cell Size

**6.Bare Nuclei** 

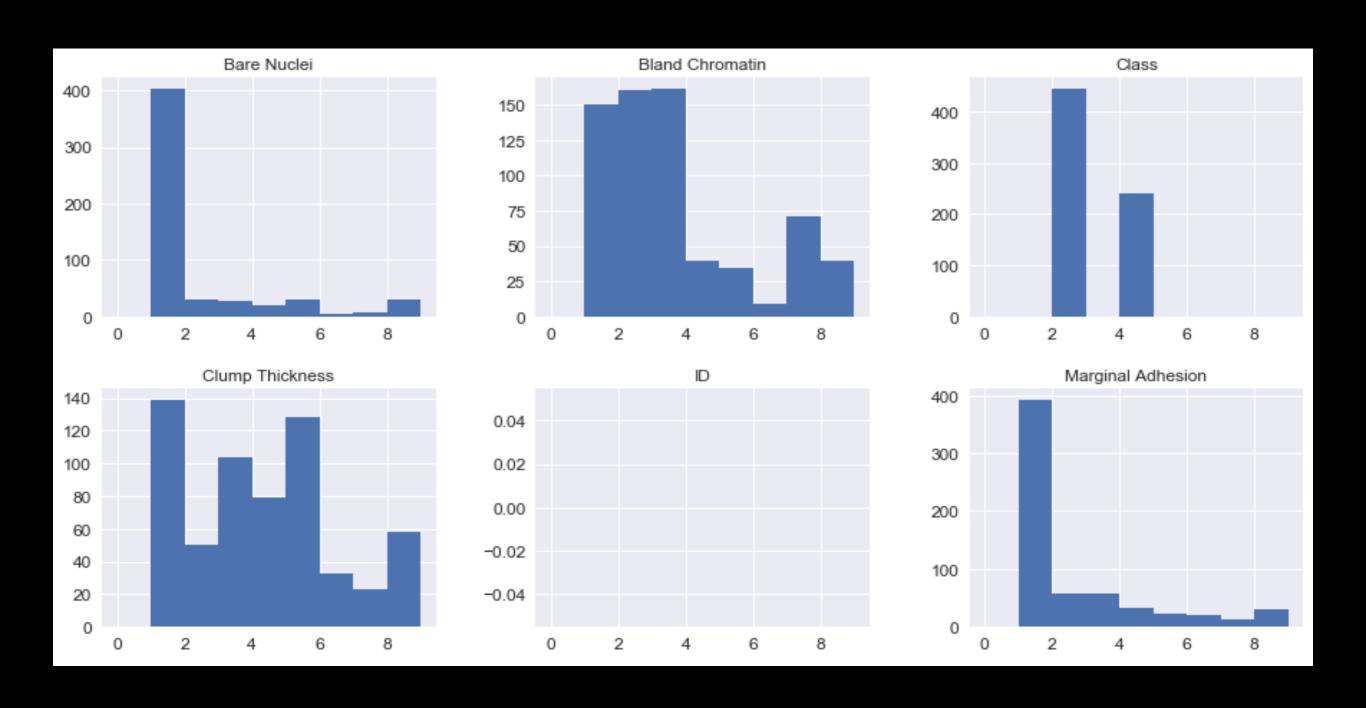
7.Bland Chromatin

**8.Normal Nucleoli** 

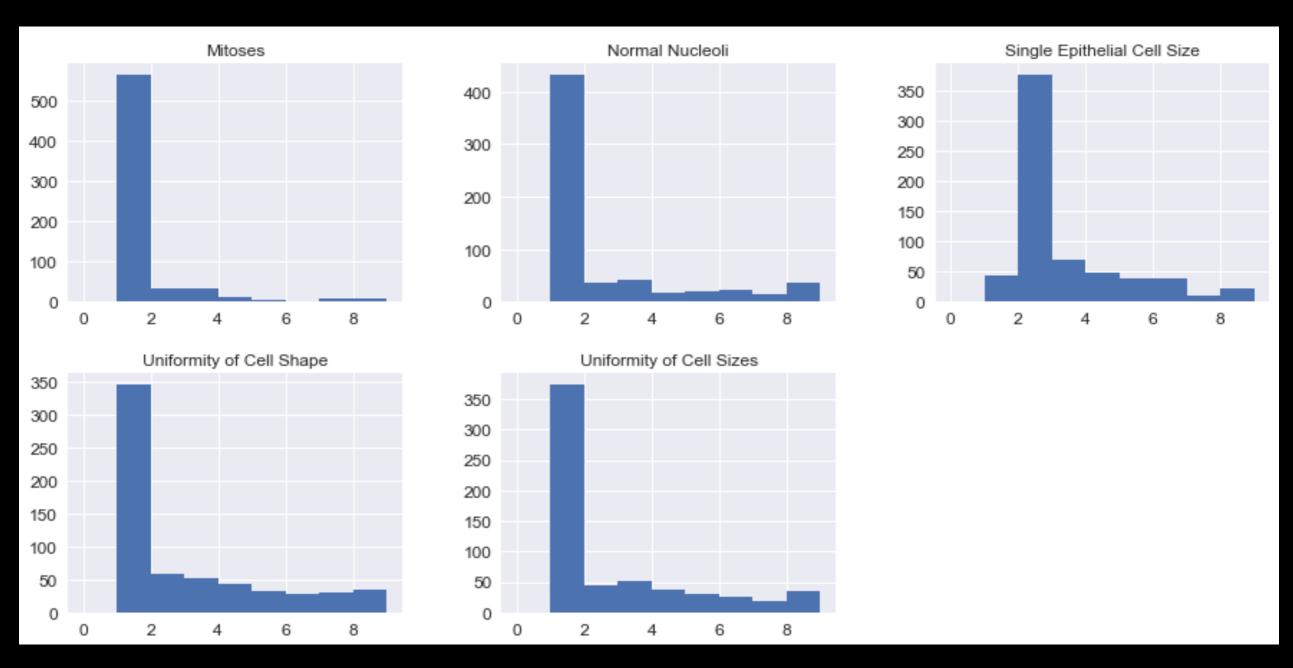
9.Mitoses

Cell Classification (i.e. Benign or Malignant)

# Input Variables

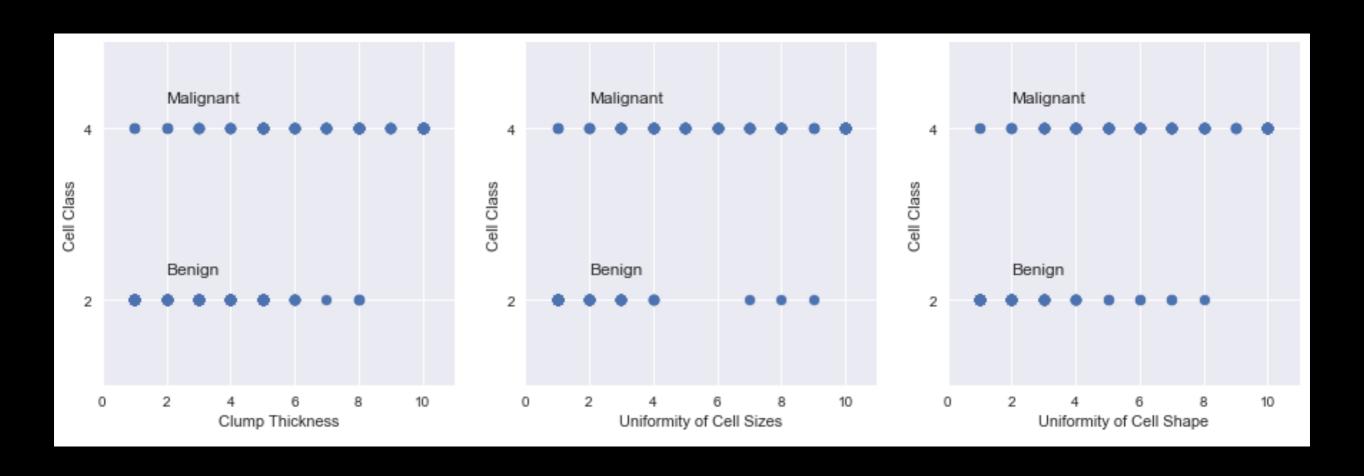


# Input Variables



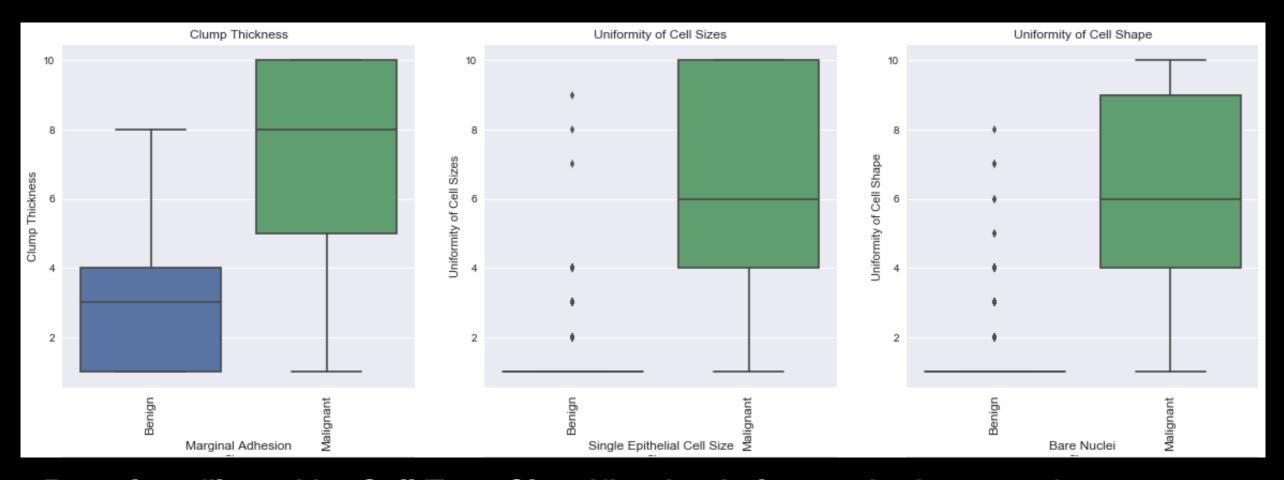
Variables are not normally distributed. They are skewed to the left. These variables clearly seem correlated.

### 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

0.75

0.60

0.45

Clump Thickness	1	0.64	0.65	0.49	0.52	0.59	0.55	0.53	0.35	0.71
Uniformity of Cell Sizes	0.64	1	0.91	0.71	0.75	0.69	0.76	0.72	0.46	0.82
Uniformity of Cell Shape	0.65	0.91	1	0.69	0.72	0.71	0.74	0.72	0.44	0.82
Marginal Adhesion	0.49	0.71	0.69	1	0.59	0.67	0.67	0.6	0.42	0.71
Single Epithelial Cell Size	0.52	0.75	0.72	0.59	1	0.59	0.62	0.63	0.48	0.69
Bare Nuclei	0.59	0.69	0.71	0.67	0.59	1	0.68	0.58	0.34	0.82
Bland Chromatin	0.55	0.76	0.74	0.67	0.62	0.68	1	0.67	0.35	0.76
Normal Nucleoli	0.53	0.72	0.72	0.6	0.63	0.58	0.67	1	0.43	0.72
Mitoses	0.35	0.46	0.44	0.42	0.48	0.34	0.35	0.43	1	0.42
Class	0.71	0.82	0.82	0.71	0.69	0.82	0.76	0.72	0.42	1
	Clump Thickness	Uniformity of Cell Sizes	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class

#### 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.