# Capstone Project 1 Project Data Storytelling

## Familiarizing with Data

The data was collected from the UCI Machine Learning website: [Link](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). The raw data was in the csv format. The following table contains the variables being studied.

|  |  |
| --- | --- |
| 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 | 1. Tumor Classification: Malignant or Benign |

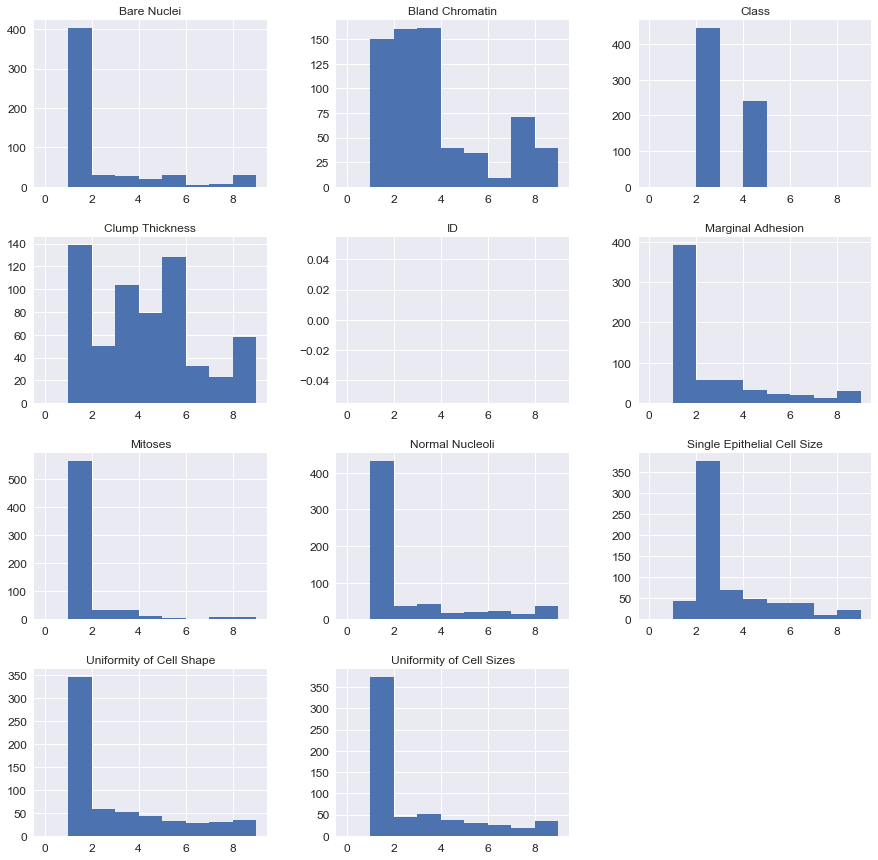
### Can you count something interesting?

65% of the data correspond to cases that have benign tumor classification. As such the data is skewed towards cases where the tumor was found to be benign.

### 

### 

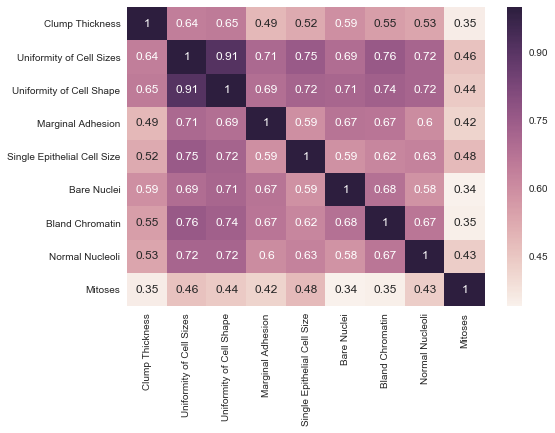
### Can you make a bar plot or a histogram? Can you find trends (e.g. high, low, increasing, decreasing, anomalies)?



Shown above is the histogram of all variables (Neglect ID variable since it has no meaning beyond identification). Following are the observations:

1. Majority of the input values are concentrated on the lower end of the 1 -10 range. This is evident in the histograms shown above.
2. Uniformity of cell shape, sizes, Marginal adhesion, Mitoses & Bare Nuclei appear to have similar profiles. Perhaps, these features are somehow correlated.

### Can you compare two related quantities?



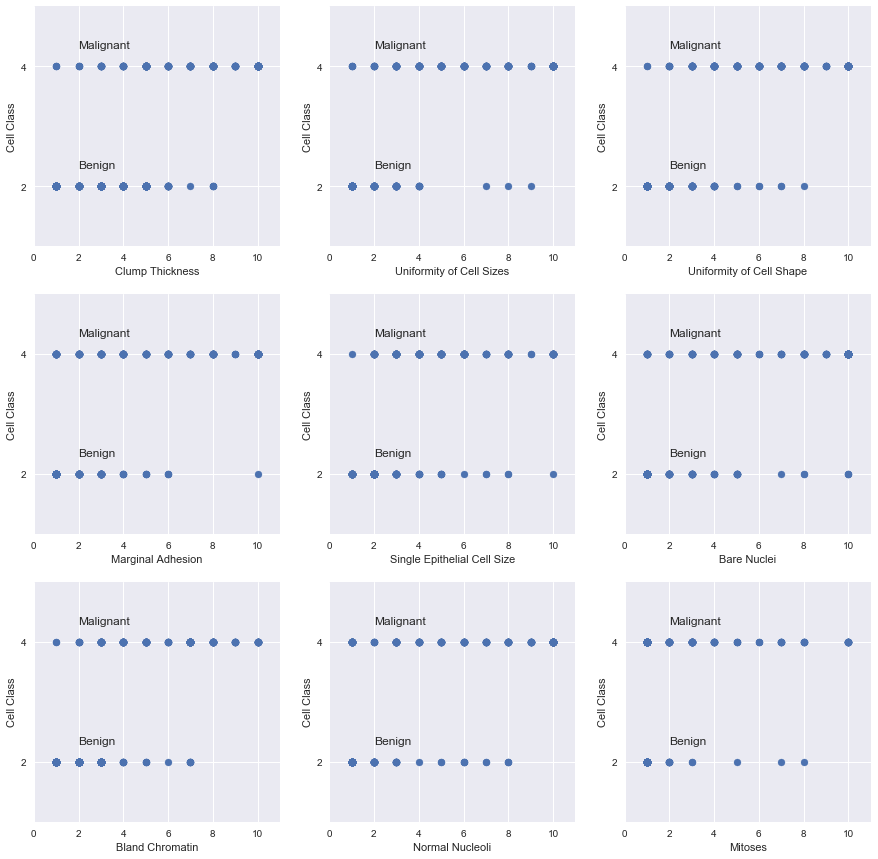
Above is the correlation matrix for all the input variables. Following can be gleaned:

1. Many of the input variables are strongly correlated with Uniformity of Cell shape & size.
2. Uniformity of cell sizes, Marginal adhesion and cell shape are highly correlated.
3. Mitoses has the least correlation with any other input variables.
4. If the goal is to build a model with the least number of input variables, then it must be built using just Uniforimity of Cell shape and Mitoses.

### 

### Can you make a scatterplot?

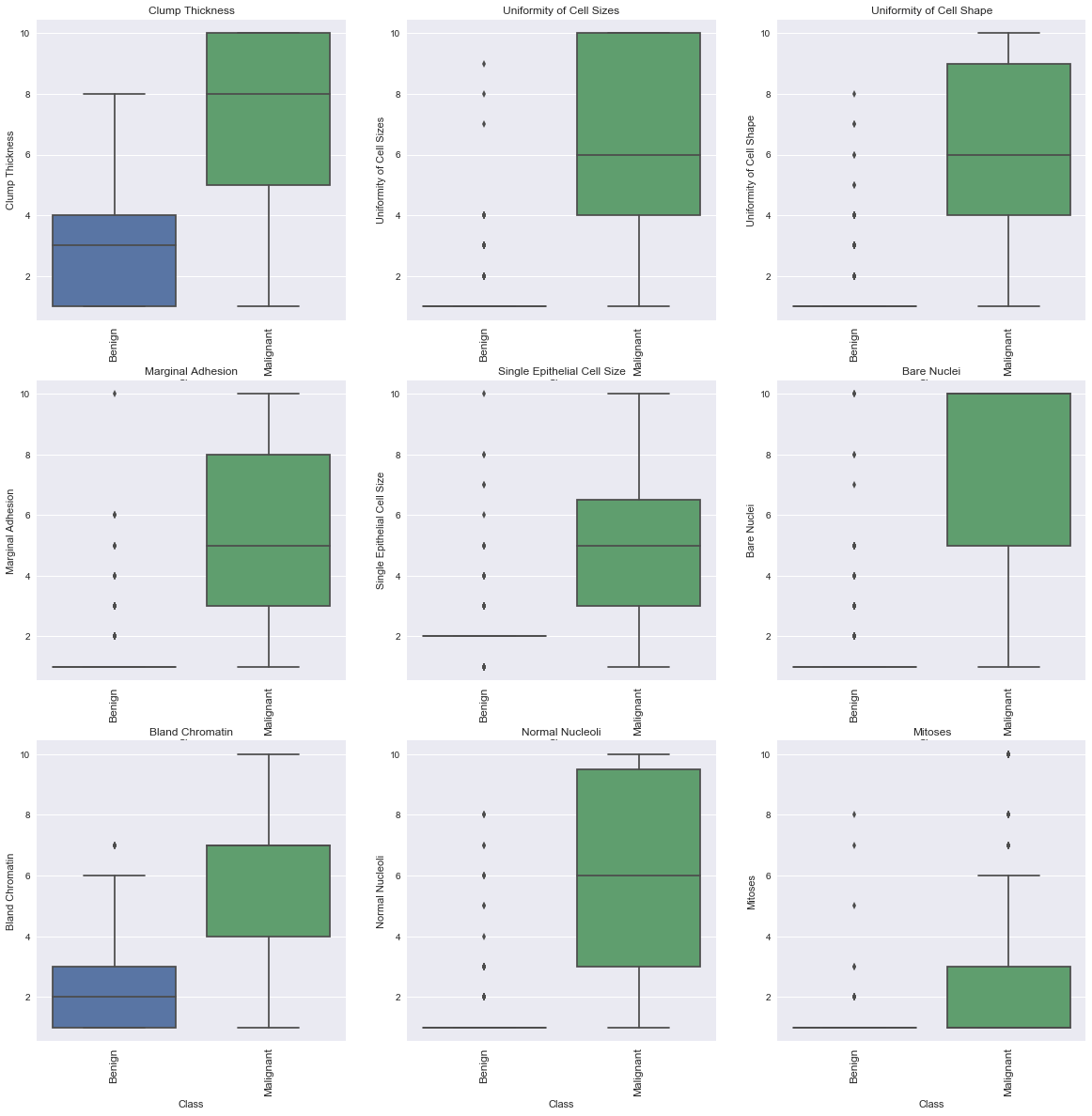
Normally scatter plots are used to identify any patterns between predictors & response variables. In this case since the response variable only takes two values and the predictors take only 9 values, the scatter plots don’t reveal much pattern. See all of the scatter plots below.



## 

## Use of Box Plots for Identifying Important Variables

Since the scatter plots don’t reveal any strong patterns between the input variables and the output variable, an alternative route was pursued. The box plots for all the variables were plotted using the cell type as the classifier. The resulting plots shows that there is fair amount of difference in the input variable values depending on whether the cell is benign or malignant. In all the cases, the value of the input variable appears to increase when the cell type changes from benign to malignant.



## Conclusions

* Most of the input variables are strongly correlated to “Uniformity of Cell size”.
* All of the input variables are heavily skewed. They do not belong to normal distribution.
* Due to the discrete nature of the variables involved, scatter plots don’t reveal any discernible patterns.
* To solve this problem box plots were prepared using the cell type as the filter. Using these it was found that the value of the input variables seems to increase when the cell type changes from benign to malignant. All of the input variables appears to have strong correlation with the cell class.