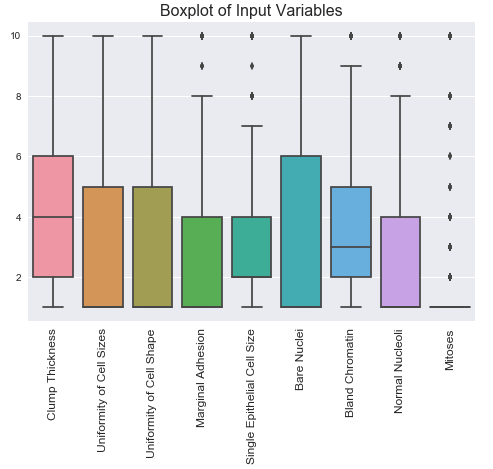
# Capstone Project 1 Project Data Wrangling

## Data Source

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. It was copied to excel and then read into a data frame.

## Data Characteristics

* There are 699 Data points, 9 input variables & one output variable. 16 missing values.
* All the input variables are qualitative in nature. They take integer values from 1 to 10.
* Excepting for cell thickness, none of them are randomly distributed. This based on the significant differences between the mean and median and the skewed nature of the population distribution as evident from the box plot.
* Owing to their non normal distribution & qualitative nature none of the data points that were past 3 standard deviations from the mean were removed.



|  |  |  |  |
| --- | --- | --- | --- |
|  | Clump Thickness | Uniformity of Cell Sizes | Uniformity of Cell Shape |
| count | 683 | 683 | 683 |
| mean | 4.4 | 3.1 | 3.2 |
| std | 2.8 | 3.0 | 2.9 |
| min | 1 | 1 | 1 |
| 25% | 2 | 1 | 1 |
| 50% | 4 | 1 | 1 |
| 75% | 6 | 5 | 5 |
| max | 10 | 10 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Marginal Adhesion | Single Epithelial Cell Size | Bare Nuclei |
| count | 683 | 683 | 683 |
| mean | 2.8 | 3.2 | 3.5 |
| std | 2.8 | 2.2 | 3.6 |
| min | 1 | 1 | 1 |
| 25% | 1 | 2 | 1 |
| 50% | 1 | 2 | 1 |
| 75% | 4 | 4 | 6 |
| max | 10 | 10 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Bland Chromatin | Normal Nucleoli | Mitoses |
| count | 683 | 683 | 683 |
| mean | 3.4 | 2.9 | 1.6 |
| std | 2.4 | 3.1 | 1.7 |
| min | 1 | 1 | 1 |
| 25% | 2 | 1 | 1 |
| 50% | 3 | 1 | 1 |
| 75% | 5 | 4 | 1 |
| max | 10 | 10 | 10 |

## Dealing with Missing Values

Only 2% of the indices had missing values. Therefore these rows were dropped from the analysis using the drop rows function in pandas.

|  |  |
| --- | --- |
| 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 |

## Correlation with Tumor Class

Boxplots of the input variables were recreated based on the tumor class. These plots show if the input variables change depending on if the tumor was benign or malignant. It is clear from these plots that the Uniformity of Cell Size, Cell Shape, Bare Nuclei, Normal Nucleoli & Mitoses vary significantly with the tumor class. These input variables should provide a good prediction of the tumor class.

