# Capstone Project 1 Exploratory Data Analysis

## Building Model Using Logistic Regression

* A logistic regression model was built using all the input variables.
* 90% of the data were used to train the model and the remaining 10% were used to test.
* Table 1 shows the confusion matrix that resulted from this model.
* The model has good accuracy (98%) and made only one mistake in classification out of 69 samples.
* Table 2 shows the coefficients of logistic regression model. Most of the variables move in the same direction as the Y variable (i.e. classification). This is also consistent with trend observed in the the box plots in Data Story Report.

Table 1. Confusion Matrix from Logistic Regression Model

|  | **Actual Benign** | **Actual Malignant** |
| --- | --- | --- |
| **Predicted Benign** | 47 | 0 |
| **Predicted Malignant** | 1 | 21 |

Table 2. Coefficients from Logistic Regression.

| **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 |

## Significant Variables

Are there variables that are particularly significant in terms of explaining the answer to your project question?

* Since all the X variables have the same range (0 to 10), the magnitude of the coefficient for each variable represents its importance.
* Based on the above principle, the top three variables were selected and a model was built based on them. The three variables are:
  + Bare Nuclei
  + Uniformity of Cell Shape
  + Clump Thickness
* The resulting model has 98% accuracy score, same as the original model built using all variables.
* Based on the above result, the three variables shown above are the significant ones. Their coefficients are shown in Table 3.

Table 3. Coefficients of Variables

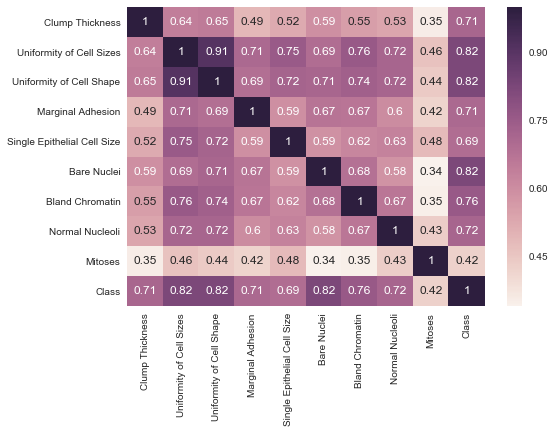
| **Variable Name** | **Coefficients** |
| --- | --- |
| **Bare Nuclei** | 0.41 |
| **Uniformity of Cell Shape** | 0.63 |
| **Clump Thickness** | 0.34 |

## Relationships between Variables

Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

This was analyzed in the Data Story Report. See below.

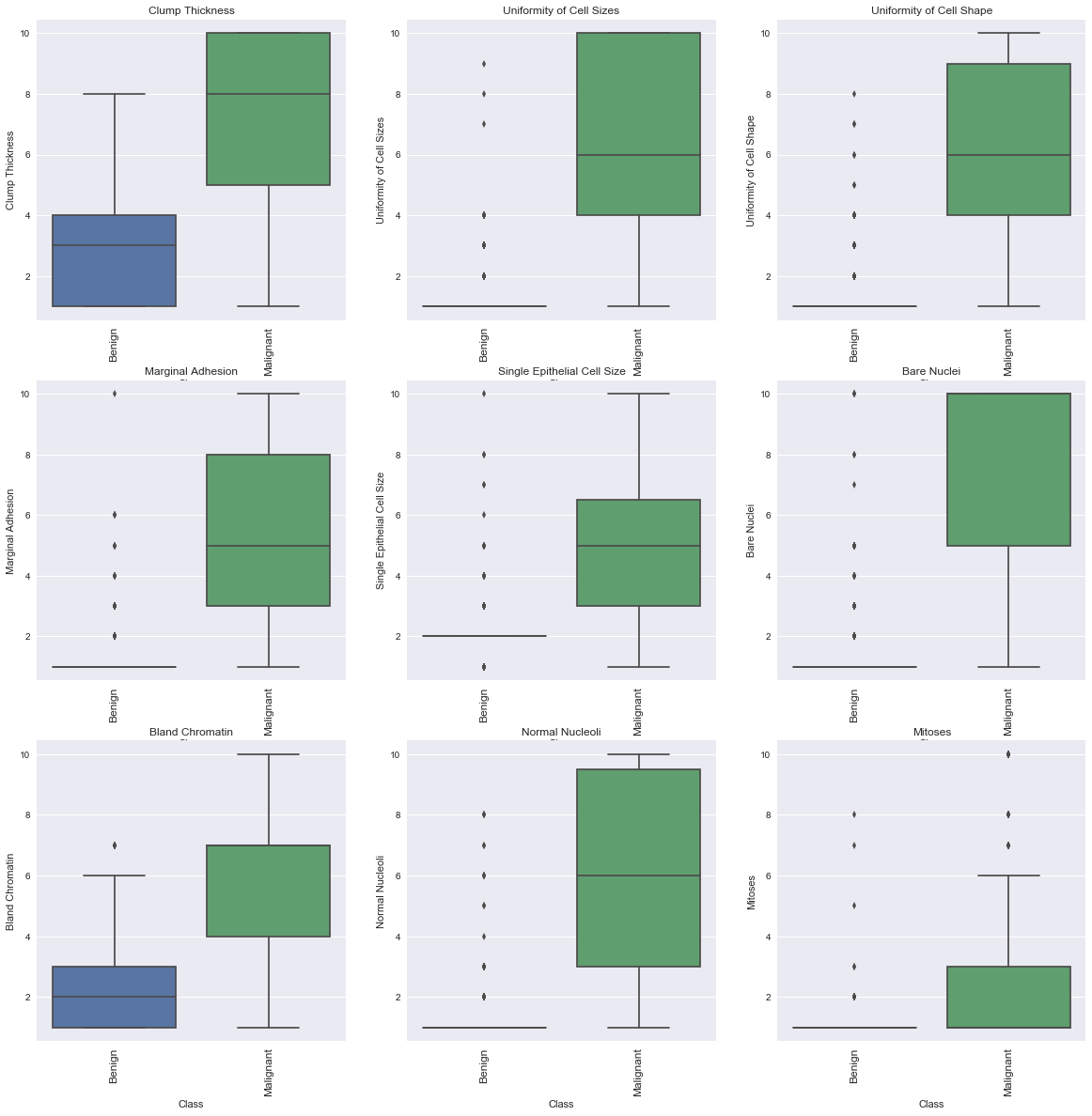
Figure 1. Correlation Matrix



Above is the correlation matrix for all the input & output 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. Excepting for Mitoses, most of the input variables have strong correlations (i.e. > 0.7) with output variables as seen in the last row. This is consistent with the trends observed in the box plot shown in Data Story Report (reproduced here as Figure 2 for convenience)
5. Cell Class (i.e. Y Variable) is most correlated with Unformity of Cell Shape/Size & Bare Nuclei. These same variables were also found to be the most significant variables from the analysis of logistic regression coefficients.

Figure 2. Box Plots Showing Relationship between Input and Output variables



## Analyzing the Relationships

What are the most appropriate tests to use to analyze these relationships?

* The confidence levels of the model accuracy and coefficients must be estimated in order to gain trust in the results.
* 1000 Runs were performed and statistics of the accuracy and model coefficients were calculated. These are shown in Table 4.
* If impact of any of the input variable on the predictor variable was uncertain, then a hypothesis test would have warranted. Here the impact of the input variables is unambiguous as seen in the correlation coefficients matrix and the box plots in the previous page. Hence no further tests are pursued.

Table 4. Results from 1000 Runs

|  | **2.5 Percentile** | **Median** | **97.5 Percentile** |
| --- | --- | --- | --- |
| **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 Classification (benign vs malignant) can be predicted with high accuracy using the input variables Bare Nuclei, Uniformity of Cell Shape & Clump thicnkess.
* A model was built using logistic regression and the three variables listed above. The model achieved > 90% accuracy in several simulated runs.