## Introduction

This report will examine the “breast\_cancer.dat” dataset, which records a series of characteristics of 90 samples of tissue from both benign and malignant masses. The report will outline the main characteristics of this dataset and its findings. The main aim of the dataset is to determine the correlation between different characteristics associated with breast tumours, such as clump thickness and uniformity of cell size. Analysis of the dataset was carried out in R, and associated graphs were also created in R.

## Methods

The variables observed in this analysis included the following:

Clump Thickness, Uniform Cell Size, Uniform Cell Shape, Marginal Adhesion, Epithelial Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class of Tumour. The first 9 variables were ranked on a standard scale of 1-10 units, and Class of Tumour was changed to values of either 0 for benign tumours or 1 for malignant tumours to allow for ease of interpretation and graphing.

Some variables were missing values and hence these were removed from the dataset to standardise the analysis. The “apply()” function was used to compute the frequencies for each value of the variables and then sort the results by tumour class. The “cor()” function was then employed to compute correlation between each pair of variables, and these values were then rounded. The “corrplot()” function was then used to visulaize the correlations between each pair of variables.

![Engineering drawing, rectangle

Description automatically generated]() *Figure 1. 1 – Findings*

## Results

The “apply()” function demonstrated that benign tumours have much lower values for all variables observed, however malignant tumours had much higher values.

The use of the “cor()” function had equally predictable results with all variables except mitoses having a positive correlation with Class of Tumour, as is evident in Figure 1.1 below. For almost all variables, with the exception perhaps of mitoses, we can infer that an increase in these values will significantly increase probability of a tumour being malignant

![A picture containing chart

Description automatically generated]()*Figure 1.2 - Colour Correlation Plot*

## Discussion

The results demonstrate that the probability of a tumour being malignant has a positive correlation with the observed Clump Thickness, Uniform Cell Size and Shape, Marginal Adhesion, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Epithelial Size, and perhaps to a lesser extent, mitoses, of the given tumour. When screening for malignancy in tumours, particular attention should be given to variables which have a strong positive correlation, which from my analysis includes Uniform Cell Size & Shape, Bare Nucleoli, and Bland Chromatin.

![Chart, box and whisker chart

Description automatically generated]() *Figure 1.3 - Findings*