MAS8404 Report

Harvey Yuan 0077439

2022-11-10

Aim and Objective

The aim of this analysis is to build a classifier for the Class (benign or malignant) of a tissue sample on some of the nine cytological characteristics of the Breast Cancer data which can be found within the mlbench library in R.

Using this data, we will hopefully be able to build a classifier for the two classes which will allow doctors and physicians to diagnose more accurately.

Technologies and Libraries Used

R version 4.1.2

R is a language used for statistical computing and graphics.

RStudio 2021.09.01 Build 372

RStudio is an IDE that is used for R. RStudio includes an in-built console and terminal that allows for direct code executing and also a pane which allows for files, plots and packages management. Different libraries can be installed through this pane.

dplyr version 1.0.7

The dplyr library allows for the user to use a set of verbs to write code for common data manipulation steps. It allows the user to use familiar words when scripting analysis and pre-processing scripts (such as filter), both for easy understanding of the code and proof reading (Hadley Wickham, n.d.).

Data Understanding and Preparation

As previously mentioned, the data used in the analysis is the BreastCancer dataset which is a part of the mlbench library. This dataset contains tissue samples from 699 women using fine needle aspiration cytology (FNAC) which extracts sames from the abnormally appearing breast tissue.

The nine cytological charateristics are stored as ordinal variables on a scale of 1-10 where the smaller the number, the healthier in term of that characteristic. As part of the nine cytological characteristics, there is also a character variable and a target class which brings the total to eleven different variables.

The eleven different variables are as below:

Id	Sample code number
Cl.thickness	Clump Thickness
Cell.size	Uniformity of Cell Size
Cell.shape	Uniformity of Cell Shape
Marg.adhesion	Marginal Adhesion
Epith.c.size	Single Epithelial Cell Size
Bare.nuclei	Bare Nuclei
Bl.cromatin	Bland Chromatin
Normal.nucleoli	Normal Nucleoli
Mitoses	Mitoses
Class	Target Class

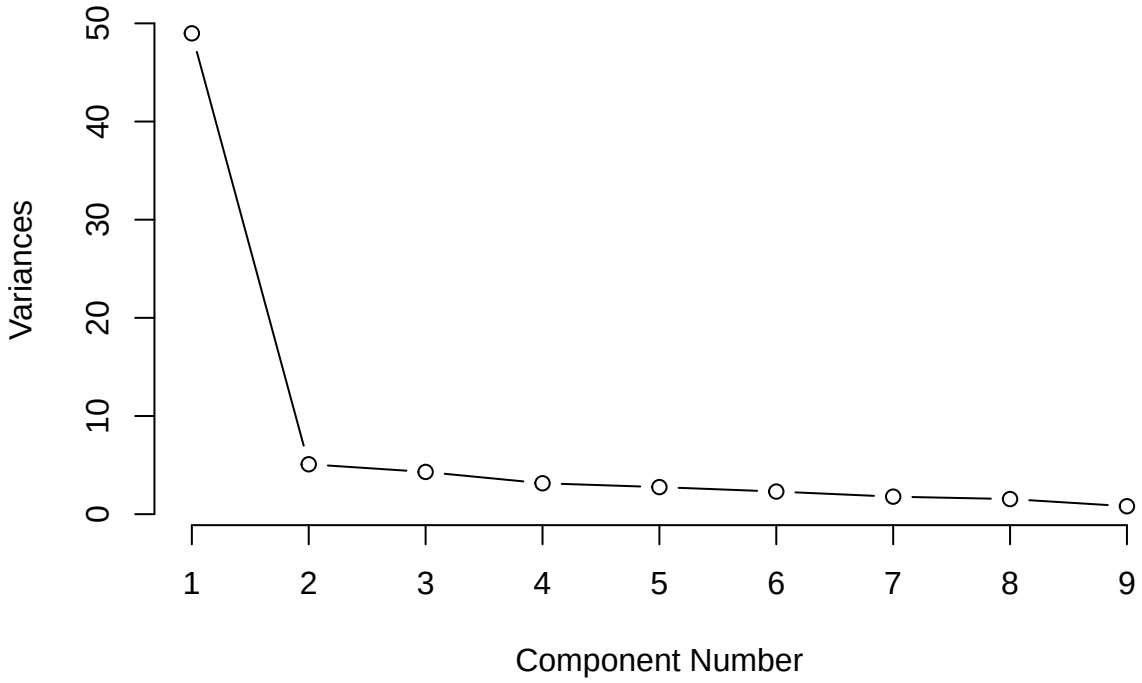
From the 699 observations, there are 16 observations which have missing observations which display as NA. For this exercise, the observations which contained NA values were ignored.

Hyun Kang describes many different techniques for handling missing data, in the paper colpublished in 2013 titled "The prevention and handling of the missing data and describes the types of missing data of whether they are missing completely at random, missing at random and missing not at random. As the data are unlikely to be missing not at random, the NA values were ignored which is outlined as listwise or case deletion method.

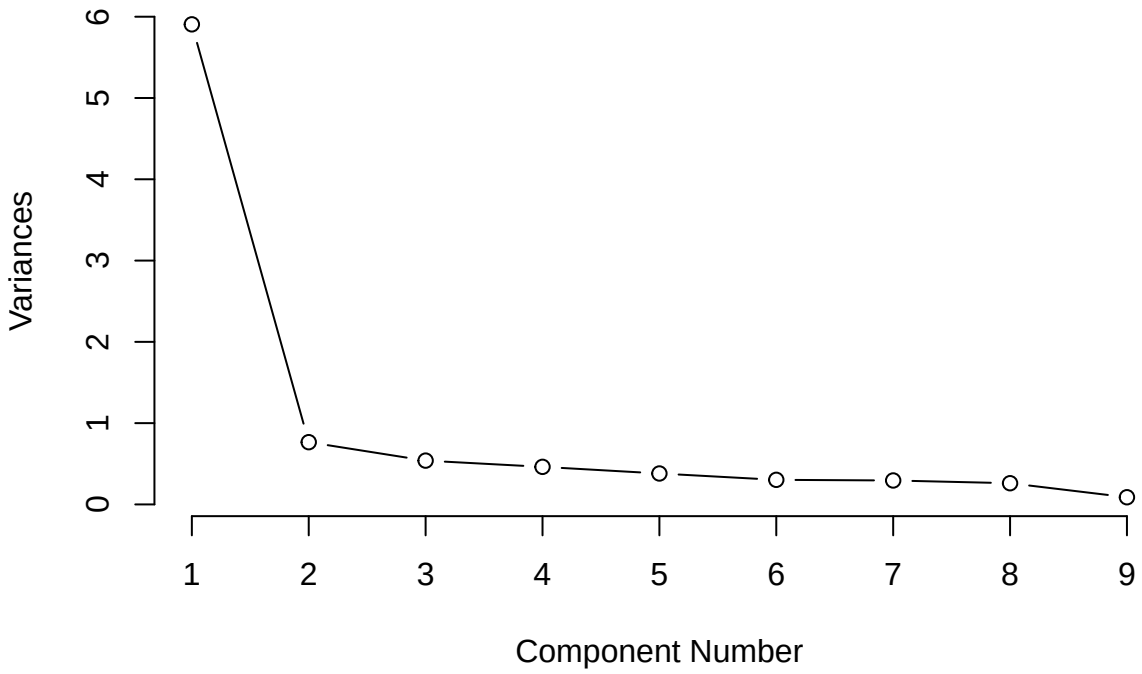
As part of the initial exploratory data analysis, the variance of the data was calculated against the no NA data. Cell size and shape ranks highly in their variance in comparison to some of the other variables. This makes sense as the size and shapes are physical traits of the cancer bodies which would explain why there's so much variation.

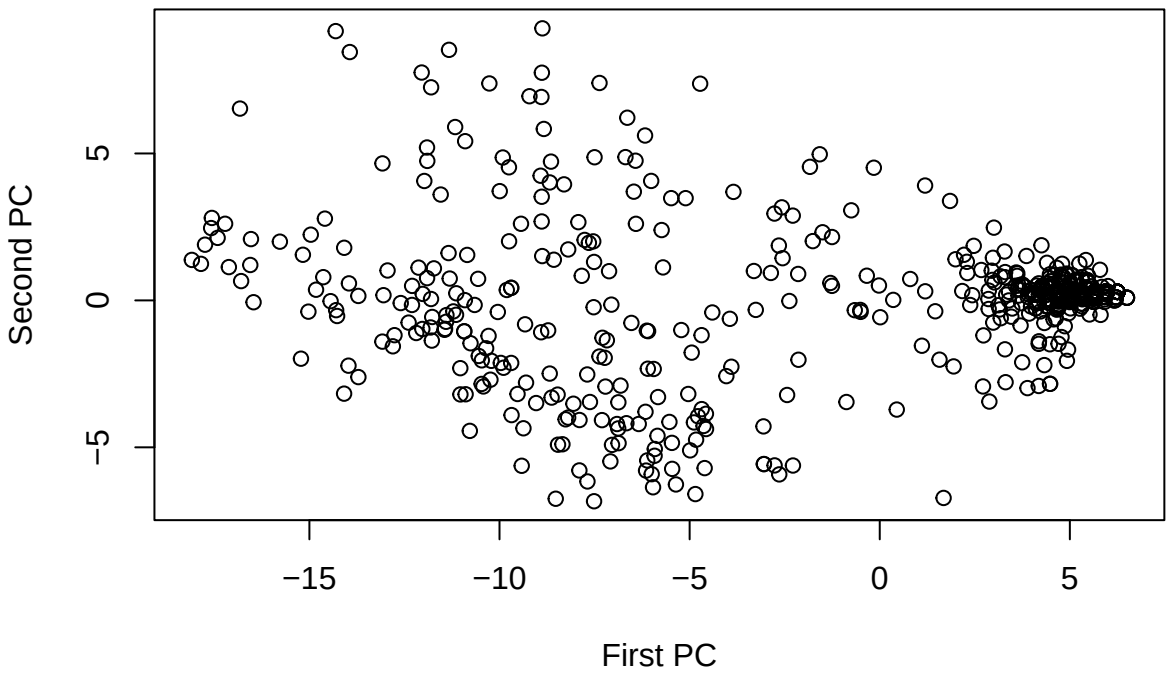
Secondly the correlation between data was also calculated which showed that the highest correlation between variables are cell size and cell shape. Once again this would make sense as one would assume that as the cell size increases, so would the cell shape. From the calculations, the cluster thickness also had a high correlation between the cell size and shape which would suggest that there is a linearity in the data between these three variables.

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     6.9988 2.25448 2.07405 1.77441 1.66043 1.51777 1.33519
## Proportion of Variance 0.6928 0.07188 0.06084 0.04453 0.03899 0.03258 0.02521
## Cumulative Proportion  0.6928 0.76466 0.82550 0.87003 0.90902 0.94160 0.96681
##                           PC8     PC9
## Standard deviation     1.24089 0.89823
## Proportion of Variance 0.02178 0.01141
## Cumulative Proportion  0.98859 1.00000
```




```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4302 0.87512 0.73417 0.67979 0.61688 0.55010 0.54274
## Proportion of Variance 0.6562 0.08509 0.05989 0.05135 0.04228 0.03362 0.03273
## Cumulative Proportion  0.6562 0.74132 0.80121 0.85256 0.89484 0.92847 0.96120
##                           PC8     PC9
## Standard deviation     0.51074 0.29730
## Proportion of Variance 0.02898 0.00982
## Cumulative Proportion  0.99018 1.00000
```


Variances
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
Component Number



For this exploration, we have chosen to reduce the dimensions of the data to where the principle components cover at least 80% of the variations. The principle component summaries above show that, for bother the raw data and the scaled data, the third principle component captures >80% of the variation.