Logistic Regression Modelling, Principal Component Analysis and Cluster Analysis
Of Breast Cancer
Dataset

LATEEF MUIZZ KOLAPO

January, 2021

Abstract

Understanding how to work with breast cancer data to aid the early detection of breast cancer in women is very important to the health and wellbeing of women around the world. This study explores various statistical methods and techniques to analyze breast cancer related dataset, to discover if common statistical methods can be used to analyze these datasets. The impact of breast cancer on the well being of women provokes the need for both accurate and interpretable results. The statistical methods investigated in this study are Logistic regression modelling, Principal component analysis and clustering analysis. Logistic regression was used on the Haberman's Breast Cancer Survival dataset to create a model representing the relationship between the variables in the dataset. We used principal component analysis to reduce the dimensionality of the anthropometric data of breast cancer patients and the control groups from the Coimbra breast cancer data. Finally, Hierarchical and K-Means clustering was used to cluster the Wisconsin breast cancer data into groups of benign and malignant breast masses.

1 Introduction

According to Global Cancer Statistics, in 2018 Lung and breast cancers were the most common cancers worldwide, each contributing 12.3% of the total number of new cases diagnosed (Bray et al., 2018). Breast cancer is considered one of the deadliest type of cancer among women, according to Cancer Research U.K, Almost half of all cancer deaths in females are from lung, breast or bowel cancer, 2017, in the UK (Cancer Research UK, 2017). Understanding how to work with breast cancer data will aid the use of statistical tools in the early prognosis of this disease in women, therefore help minimize its lasting and scaring effect on the patients. Large amounts of breast cancer data have been gathered and generated by different research and medical institutes over the years such as anthropometric data of the patients, surgery data of the patients and data on physical attributes of the breast masses. The continuous advancement in the field of big data, data science and statistics coupled with cheaper cost of computing power means a wide range of techniques are now available to be used to understand this data, three of which were applied in this study (Sivarajah et al., 2017).

2 Methods and Methodology

Three analysis were performed using three different datasets including Logistic Regression, Principal Component Analysis and Clustering Analysis. The SAS software environment was used to perform the three tasks.

2.1 Research 1: Design a model reperesenting the odds of a patient surviving more than 5 years after a breast cancer surgery using Age of patient at time of operation and Number of positive axillary nodes detected.

Dataset Description: The dataset used for the implementation of the Logistic Regression model is called the Haberman's Survival Data which was sourced here. The dataset describes the survival rates of breast cancer patients who had undergone surgery between the year 1958 and 1970. The dataset is a multivariate dataset with 306 observations and 4 attributes illustrated in Figure 1. Each row in this dataset represents the surgery details of a patient that survived 5 years or longer and patients that died within 5 years.

List of Variables and Attributes		
Variable	**Type**	**Description**
Age	Num	Age of patient at time of operation.
Number of Axillary Node	Num	Number of positive axillary nodes detected
Survival status	Num	Class attribute. 1 = the patient survived 5 years or longer and 2 = the patient died within 5 year
Year of operation(1900)	Num	Patient's year of operation.

Figure 1: Feature Information For the Haberman's Survival Data.

Variables Used: Logistic regression was used to model the odds of a patient surviving 5 years or longer using Age of patient at time of operation and Number of positive axillary nodes detected.

Binary Outcome	
Class	**Class Information**
1	Patient survived 5 years or longer.
2	Patient died within 5 years.

Table 1: Table of Outcomes

Model Information

Figure 2 shows details about the dataset and the model. Fisher's scoring algorithm was used to optimize the maximum likelihood for the model, all observations in the data were used. The model converged using gradient convergence criterion(GCONV) with a precision of 10^{-8}.

Model Information		
Data Set	CW1_4S08.HABERMAN_REGRESSION	
Response Variable	Survival status	Survival status
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	306
Number of Observations Used	306

Response Profile		
Ordered Value	**Survival status**	**Total Frequency**
1	1	225
2	2	81

Probability modeled is Survival status='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Figure 2: Model Information

Model Fit

The model fit statistics explain the overall fit of our model, it inform us if having an Intercept only is better than having Intercept and Covariates. The Akaikae Information(AIC) and Schwarz Criterion in Figure 3 were used in this study to

explain the overall fit of the model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	355.688	334.311
SC	359.412	345.481
-2 Log L	353.688	328.311

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.3775	2	<.0001
Score	27.4399	2	<.0001
Wald	20.9934	2	<.0001

Figure 3: Model Fit Statistics and Global Null Hypothesis Test.

Definition of Model Fit Statistics	
Criterion	**Definition**
AIC	This is the Akaike Information Criterion, calculated as $AIC = -2LogL + 2((k-1)+s)$, where k is the number of the dependent variable in the model and s is the number of predictor variables in the model. The model with smallest AIC is considered the best option. AIC favors a simpler model over a complicated model and punishes our model based on the number of independent variables present. The Intercept and Covariates was selected as the best fit for this model since it has a smaller value 334.311 than the intercept 355.688.
SC	It is defined as ˘$2LogL + ((k-1)+s) * Log(\sum fi)$, where fi‘s are the frequency values of the ith observation, k and s are the same as was defined above. SC punishes for the number of independent variables and observations in the model and the smallest SC is most preferred. The Intercept and Covariates are the best option for this model since we observed a value 345.481 smaller than the observed value of the intercept 359.412.

Table 2: Model Fit Statistics Definitions

Based on the evidence above AIC and SC confirms the variables will improve our model.

Testing Global Null Hypothesis

Hypothesis and Assumption	
Hypothesis	**Assumption**
H_0	All the predictor variable in this model are not significant i.e $b_1 = b_2 = 0$.
H_1	At least one of the predicor variable in this model is significant i.e $b_1 or b_2 \neq 0$.

Table 3: Global Null Hypothesis Assumptions

We performed 3 tests and their respective chi-square values are shown in Figure 3. The degrees of freedom is two since variables Age and Number of positive axillary nodes detected are being used in this study. At a significance value of 0.05, the observed p-value is lower, this means we have enough evidence to reject the Null hypothesis that the predictor is not significant. We can confirm our model is better than an empty model.

Analysis of Maximum Likelihood Estimates

Figure 4 provides information on the coefficients for the parameters. These are the values for the parameters β_1 and β_2 in our logistic regression equation.

$$log(odds) = log(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The DF column in Figure 4 shows the Degree of freedom which is one for each variable since this test considers individual variables. The value of the parameters β_1 and β_2 are observed in the Estimate column. The standard errors of the individual regression coefficients was recorded under standard error. The Wald chi-square, and the P-value which are the test statistics and p-values, respectively, were used to test the null hypothesis that an individual predictor's regression coefficient is zero, given the other predictor variables are present in the model.

The hypothesis is as seen below:

Hypothesis and Assumption	
Hypothesis	**Assumption**
H_0	estimates = 0 for the individual predictor i.e the marginal contribution of the variable given the other variables are present in the model is zero.
H_1	estimates $\neq$ 0 for the individual predictor i.e the marginal contribution of the variable given the other variables are present in the model is not zero.

Table 4: Analysis of Maximum Likelihood Hypothesis Assumptions

Figure 4 shows that at a significance level of 0.05, we do not have enough evidence to reject the null hypothesis for the Age variable. The output reveals we have enough evidence to reject the null hypothesis for the number of positive axillary nodes used. This estimates can be interpreted as — For one unit change in the number of positive axillary nodes detected, the difference in the log-odds for surviving more than 5 years is expected to change by 0.0884 given other variables in the model are held constant.
The logistic regression model for this relationship can then be expressed with the equation below:

$$Logit(P) = 2.4629 - 0.0197 * Age - 0.0883 * \text{number of positive axillary nodes detected}$$

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.4629	0.7064	12.1551	0.0005
Age	1	-0.0197	0.0127	2.3987	0.1214
Number of positive a	1	-0.0883	0.0198	19.8586	<.0001

Figure 4: Analysis of Maximum Likelihood Estimates

Odds Ratio Estimate

The odds ratio estimate in Figure 5 gives the coefficient of the odds ratio for the predictor variables which is the exponentiated parameter for the predictor e^b. The 95% Wald confidence limits indicates that for any of the predictors in the model, we are 95% confident that if the experiment was repeated, we can expect $\approx$ 95% of the confidence intervals to include our point estimate value. Our odds ratio can be interpreted as follows:

- A unit change in the number of positive axillary nodes detected changes the odds of survival by 0.915, given other variables in the model are held constant.
- A unit change in the age of the patient changes the odds of survival by 0.980, given other variables in the model are held constant.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.981	0.956	1.005
Number of positive a	0.915	0.881	0.952

Figure 5: Table for Odds Ratio Estimates.

Association of Predicted Probabilities and Observed Responses

Figure 6 summarizes the ability of our model to discriminate between survivors and non-survivors. Concordance statistic c indicates that 70% of the time, our model is able to correctly sort a survival and non-survival pair correctly. Figure 7 shows that the probability of surviving more than 5 years after the operation has a downward trend as the age increases.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.1	Somers' D	0.407
Percent Discordant	29.5	Gamma	0.408
Percent Tied	0.4	Tau-a	0.159
Pairs	18225	c	0.703

Figure 6: Predicted Probabilities and Observed Responses

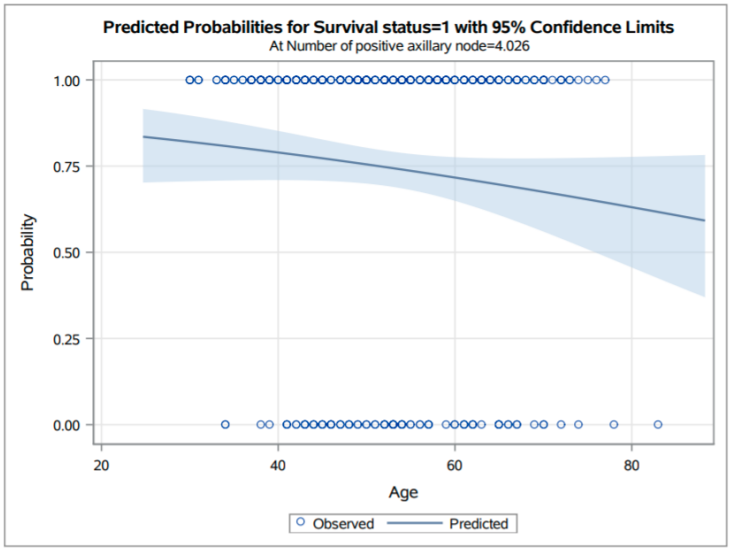

Figure 7: Plot for Predicted Probabilities for Survival

2.2 Research 2: Reduce the Dimensionality of the Coimbra Breast Cancer Data using Principal Component Analysis.

Dataset Description: The dataset used for performing the principal component analysis in this study is the Breast Cancer Coimbra Data Set found here from the Faculty of Medicine of the University of Coimbra. This is a high dimension data and principal component analysis was used to reduce the dimensionality of this dataset.

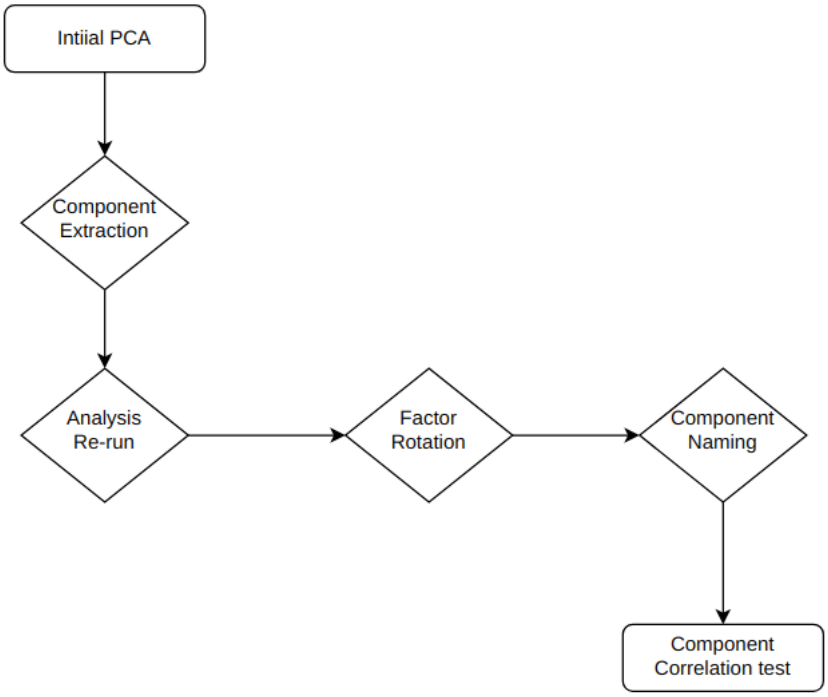

Figure 8: Architecture of Principal Component Analysis.

Variables Used: The dataset has 116 observations and 10 variables gotten from 64 patients with breast cancer and a control group with 52 patients has seen in Figure 9.

List of Variables and Attributes		
Variable	**Type**	**Description**
Age	Num	Age of patient.
BMI	Num	Body Mass Index measured in Kg/m2.
Glucose	Num	Patient glucose level measured in mg/dL.
Insulin	Num	Patient insulin level measured in µU/mL
Homa	Num	Homeostatic model assessment
Leptin	Num	leptin measured in ng/mL
Adiponectin	Num	Adiponectin measured in µg/mL
Resistin	Num	Resistin measured in ng/mL
MCP.1	Num	monocyte-chemoattractant protein-1 (MCP-1/CCL2) measured in pg/dL
Classification	Char	Class attribute. 1=Healthy controls, 2=Patients (with cancer)

Figure 9: Feature Information For Breast Cancer Anthropometric Data.

Exploratory Analysis

The table in Figure 10 shows details of the dataset used for this analysis. The software used all the instances which means we have no missing instances in the dataset.

Input Data Type	Raw Data
Number of Records Read	116
Number of Records Used	116
N for Significance Tests	116

Means and Standard Deviations from 116 Observations		
Variable	**Mean**	**Std Dev**
Age	57.30172	16.11277
BMI	27.58211	5.02014
Glucose	97.79310	22.52516
Insulin	10.01209	10.06777
HOMA	2.69499	3.64204
Leptin	26.61508	19.18329
Adiponectin	10.18087	6.84334
Resistin	14.72597	12.39065
MCP.1	534.64700	345.91266

Figure 10: Summary Statistics of Dataset .

The Pearson's Correlation Coefficient and Scatter Plot Matrix in Figure 11 and 12 respectively reveals a strong correlation between Homa and Insuline. We would be leaving them in the analysis because they offer different explanations to our dataset according to definitions from (Wallace et al., 2004) and (Masoud et al., 2006).

9 Variables:	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1

Pearson Correlation Coefficients, N = 116
Prob > |r| under H0: Rho=0

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Age	1.00000	0.00853 0.9276	0.23011 0.0130	0.03250 0.7291	0.12703 0.1742	0.10263 0.2730	-0.21981 0.0177	0.00274 0.9767	0.01346 0.8860
BMI	0.00853 0.9276	1.00000	0.13885 0.1372	0.14530 0.1197	0.11448 0.2211	0.56959 <.0001	-0.30273 0.0010	0.19535 0.0356	0.22404 0.0156
Glucose	0.23011 0.0130	0.13885 0.1372	1.00000	0.50465 <.0001	0.69621 <.0001	0.30508 0.0009	-0.12212 0.1916	0.29133 0.0015	0.26488 0.0041
Insulin	0.03250 0.7291	0.14530 0.1197	0.50465 <.0001	1.00000	0.93220 <.0001	0.30146 0.0010	-0.03130 0.7388	0.14673 0.1160	0.17436 0.0612
HOMA	0.12703 0.1742	0.11448 0.2211	0.69621 <.0001	0.93220 <.0001	1.00000	0.32721 0.0003	-0.05634 0.5481	0.23110 0.0126	0.25953 0.0049
Leptin	0.10263 0.2730	0.56959 <.0001	0.30508 0.0009	0.30146 0.0010	0.32721 0.0003	1.00000	-0.09539 0.3084	0.25623 0.0055	0.01401 0.8814
Adiponectin	-0.21981 0.0177	-0.30273 0.0010	-0.12212 0.1916	-0.03130 0.7388	-0.05634 0.5481	-0.09539 0.3084	1.00000	-0.25236 0.0063	-0.20069 0.0308
Resistin	0.00274 0.9767	0.19535 0.0356	0.29133 0.0015	0.14673 0.1160	0.23110 0.0126	0.25623 0.0055	-0.25236 0.0063	1.00000	0.36647 <.0001
MCP.1	0.01346 0.8860	0.22404 0.0156	0.26488 0.0041	0.17436 0.0612	0.25953 0.0049	0.01401 0.8814	-0.20069 0.0308	0.36647 <.0001	1.00000

Figure 11: Pearsons Correlation Coefficient Matrix of Dataset.

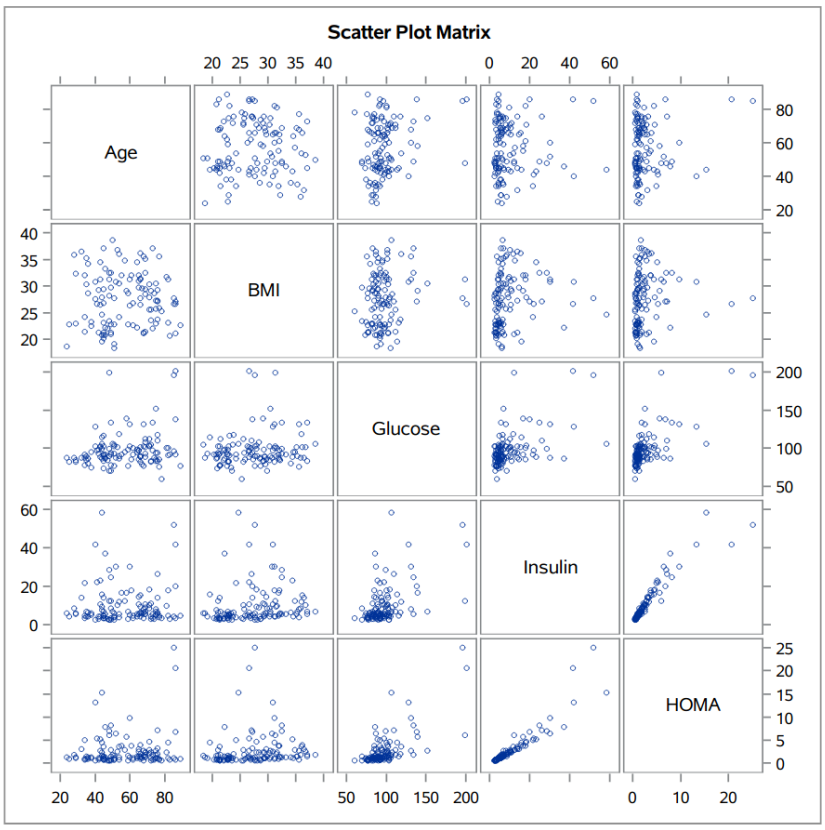

Figure 12: Scatterplot Matrix of Variables in the Dataset.

Principal Components and Eigenvalues

Figure 13 displays the Eigenvalue for the different components, the difference column shows the difference between the current Eigenvalue and the next one which shows us the changes in the values from one component to the next. The proportion column gives us the percentage of the variance explained by each of the component while cumulative does a rolling sum of the proportion. Component extraction was done using 3 techniques:

- Kaisers Rule: Kaiser (1959) recommended we only include components in the analysis with an Eigenvalue greater

than one.

- Proportion of Variance: The number of components to be retained in the analysis can be decided by choosing the number of components that account for a pre-defined amount of variation in the dataset.

$$\% \text{ of variation} = (\sum_{i=1}^{k} \lambda_i)/m$$

- Scree Plots: The Scree plot graphs show the plot of the Eigenvalue against the component number. We would be selecting components with Eignenvalues above the point where the values start decreasing linearly.

We would be considering the 3 proposed methods above.
Kaiser's Rule: Figure 13 shows the top four components have Eigenvalues greater than 1. Based on Kaiser's rule components 1 to 4 in our dataset would be extracted.

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.05853968	1.53633845	0.3398	0.3398
2	1.52220124	0.35465743	0.1691	0.5090
3	1.16754381	0.06184570	0.1297	0.6387
4	1.10569811	0.38315576	0.1229	0.7616
5	0.72254235	0.06526318	0.0803	0.8418
6	0.65727917	0.21572867	0.0730	0.9149
7	0.44155050	0.14892657	0.0491	0.9639
8	0.29262393	0.26060271	0.0325	0.9964
9	0.03202122		0.0036	1.0000

Figure 13: Eigenvalues for the Compoonents.

Proportion of Variance by Components: We would extract number of components accounting for at least 70% of the variance in the dataset. In Figure 13 we can observe that components 1 to 4 accounts for $\approx 76\%$ of the variance in our dataset which is 6% more than the predefined limit.

Scree Plot: Combining Figure 14 with the Kaisers rule we see that components with Eigenvalues lower than one in the plot are declining in a linear pattern which shows us that components with Eigenvalues below one generally account for lesser variance than the original variable.

Based on the evidences above, using the three methods for component extraction, we would keep only components with Eigenvalues greater than 1.

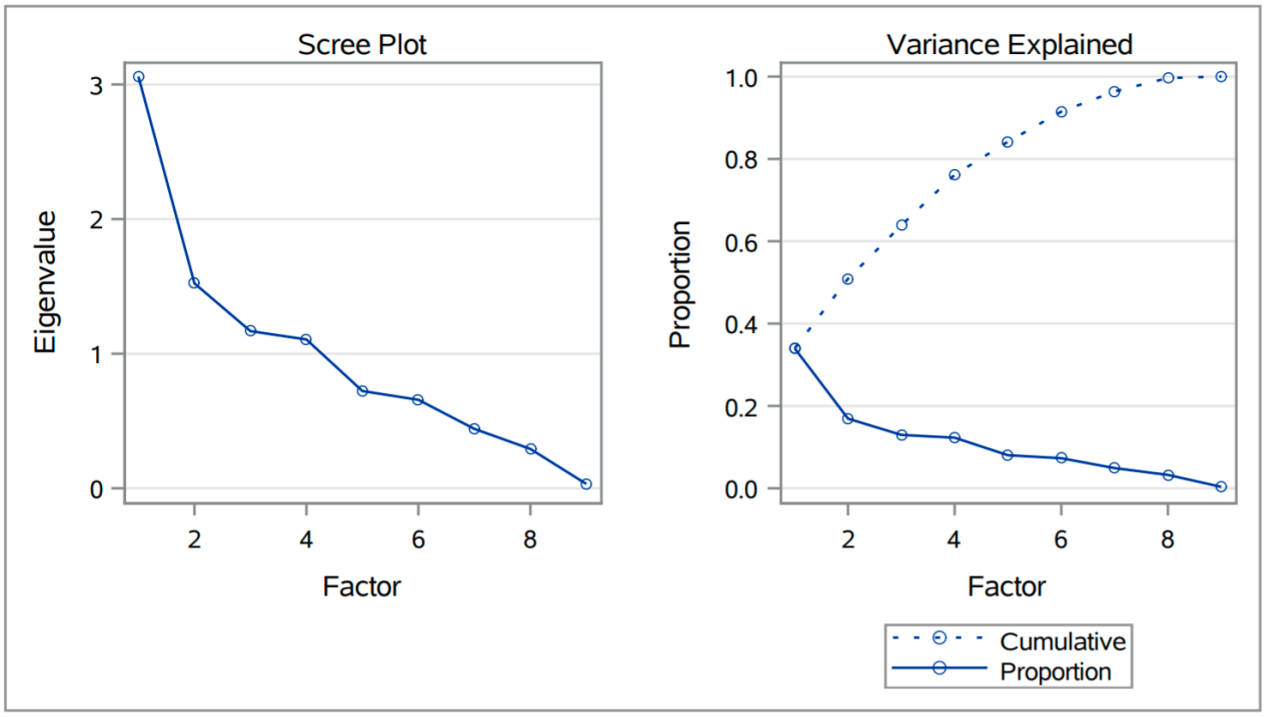

Figure 14: Scree Plot of Eigenvalues and Components

Analysis rerun

The initial analysis showed that we would be keeping the four components. Principal Component Analysis was repeated while keeping just four components.

Factor Pattern

Figure 15 shows the Component loadings which explains the correlations between each variable and the components explained in Figure 16.

Factor Pattern					
		Factor1	**Factor2**	**Factor3**	**Factor4**
Age	Age	0.21786	-0.08175	-0.22335	-0.86371
BMI	BMI	0.45546	-0.61607	0.46002	0.07458
Glucose	Glucose	0.76779	0.22941	-0.14143	-0.13209
Insulin	Insulin	0.77646	0.47662	0.10126	0.06285
HOMA	HOMA	0.86193	0.46233	-0.01317	0.00593
Leptin	Leptin	0.57974	-0.28826	0.63017	-0.06136
Adiponectin	Adiponectin	-0.30187	0.59295	0.30484	0.29113
Resistin	Resistin	0.49273	-0.37460	-0.31219	0.31830
MCP.1	MCP.1	0.44532	-0.25965	-0.53676	0.37799

Figure 15: Factor Pattern

Factor Pattern interpretation	
Variable	**Factor correlation**
Age	Age has a strong negative correlation with Factor 4.
BMI	BMI has a strong negatove correlation with Factor 2.
Glucose	Glucose has a strong positive correlation with Factor 1.
Insulin	Insulin has a strong positive correlation with Factor 1.
Homa	has a strong positive correlation with Factor 1.
Leptin	leptin has a strong positive correlation with Factor 3 and weak positive correlation with Factor 2.
Adiponectin	Adiponectin does not seem to have any strong correlation, it has a weak positive correlation with Factor 2,3 and weak negative correlation with Factor 1.
Resistin	Resistin has a weak positive correlation with Factor 1, 4 and weak negative correlation with Factor 2, 3.
MCP.1	MCP.1 has a weak positive correlation with Factor 1 and Factor 4 and a weak negative correlation with Factor 3.

Figure 16: Factor Pattern Interpretation.

Variance by Individual Factors

Variance explained by each of the factors in Figure 17 on summation gives the approximate value of 6.9, this means the 4 factors explain approximately 6.9 of the 9 variance explained by the original variable. The final communality estimates in Figure 17 equals the summation of the variance explained by the factors and shows individual contribution of the variables.

Variance Explained by Each Factor			
Factor1	**Factor2**	**Factor3**	**Factor4**
3.0585397	1.5222012	1.1675438	1.1056981

Final Communality Estimates: Total = 6.853983								
Age	**BMI**	**Glucose**	**Insulin**	**HOMA**	**Leptin**	**Adiponectin**	**Resistin**	**MCP.1**
0.85002143	0.80416471	0.67958311	0.84426457	0.95688562	0.82006666	0.62040267	0.58188179	0.69671226

Figure 17: Variance Explained by Each Factor and Communality Estimates.

Rotation

Varimax rotation which is a type of Orthogonal rotation which maximises the variance of the loadings within the principal components across the variables was used to maximise the high correlations and minimise low correlations within our dataset.

Factor Pattern

Figure 18 shows the result of our varimax rotation which has maximised and minimised correlations betwen the variables and the components interpretetd in Figure 19.

Rotated Factor Pattern					
		Factor1	**Factor2**	**Factor3**	**Factor4**
Age	Age	0.15045	-0.03089	-0.12283	0.90075
BMI	BMI	-0.00810	0.86526	0.23145	0.04309
Glucose	Glucose	0.75158	0.08883	0.22707	0.23508
Insulin	Insulin	0.90496	0.12887	0.04583	-0.08131
HOMA	HOMA	0.96457	0.09020	0.13330	0.02425
Leptin	Leptin	0.29650	0.85333	-0.04677	0.04233
Adiponectin	Adiponectin	0.11392	-0.23820	-0.47147	-0.57306
Resistin	Resistin	0.14875	0.19547	0.72216	0.00571
MCP.1	MCP.1	0.17561	-0.04880	0.81435	-0.01804

Figure 18: Table for Rotated Factor Pattern.

Factor Pattern interpretation	
Variable	**Factor correlation**
Age	Age has a strong positive correlation with Factor 4.
BMI	BMI has a strong positive correlation with Factor 2.
Glucose	Glucose has a strong positive correlation with Factor 1.
Insulin	Insulin has a strong positive correlation with Factor 1.
Homa	HOMA has a strong positive correlation with Factor 1.
Leptin	Leptin has a strong positive correlation with Factor 2.
Adiponectin	Adiponectin has a weak positive correlation with Factor 3 and 4.
Resistin	Resistin has a strong positive correlation with Factor 3.
MCP.1	MCP.1 has a strong positive correlation with Factor 3.

Figure 19: Interpretation for Rotated Factor Pattern.

Figure 20 reveals a more balanced distribution of variance explained by each of the factors.

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
2.4907588	1.6077713	1.5492327	1.2062200

Final Communality Estimates: Total = 6.853983								
Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
0.85002143	0.80416471	0.67958311	0.84426457	0.95688562	0.82006666	0.62040267	0.58188179	0.69671226

Figure 20: Variance Explained by Each Factor After Rotation.

The factor pattern diagram in Figure 21 shows the variables before and after the rotation.

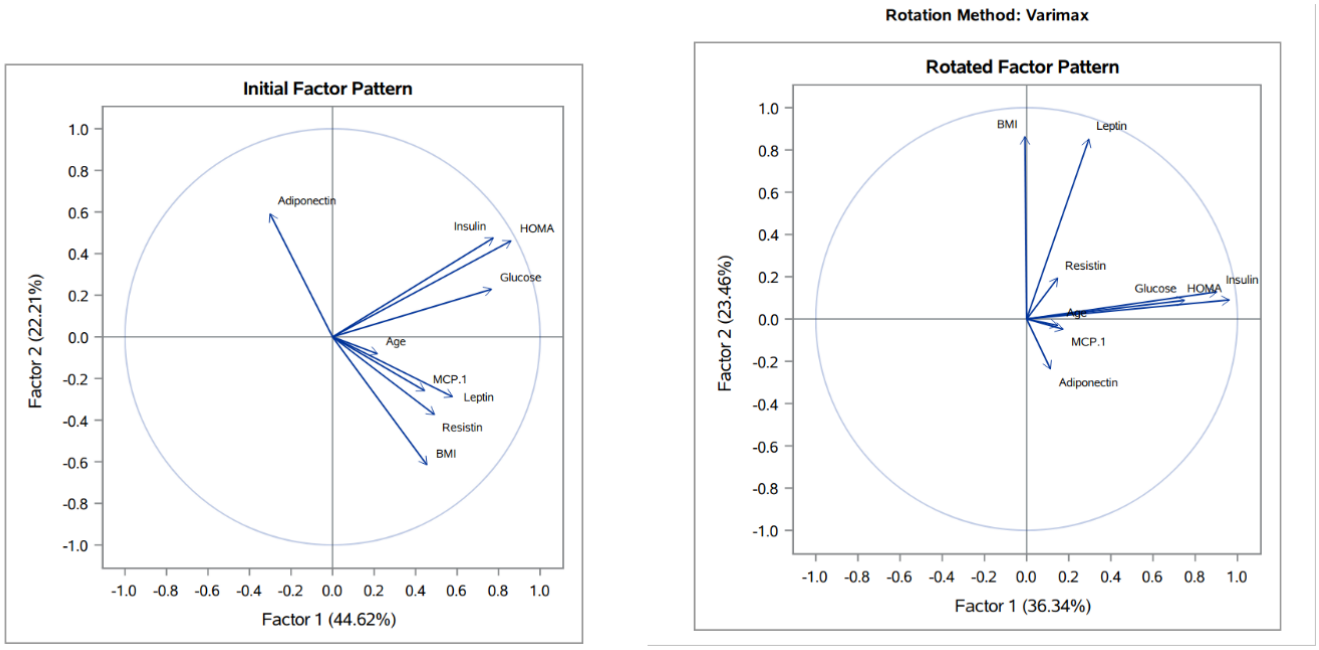

Figure 21: Rotated and Non-rotated Factor Pattern.

The path diagram in Figure 22 also shows the amount of variation accounted for in each measurement by the new Factors.

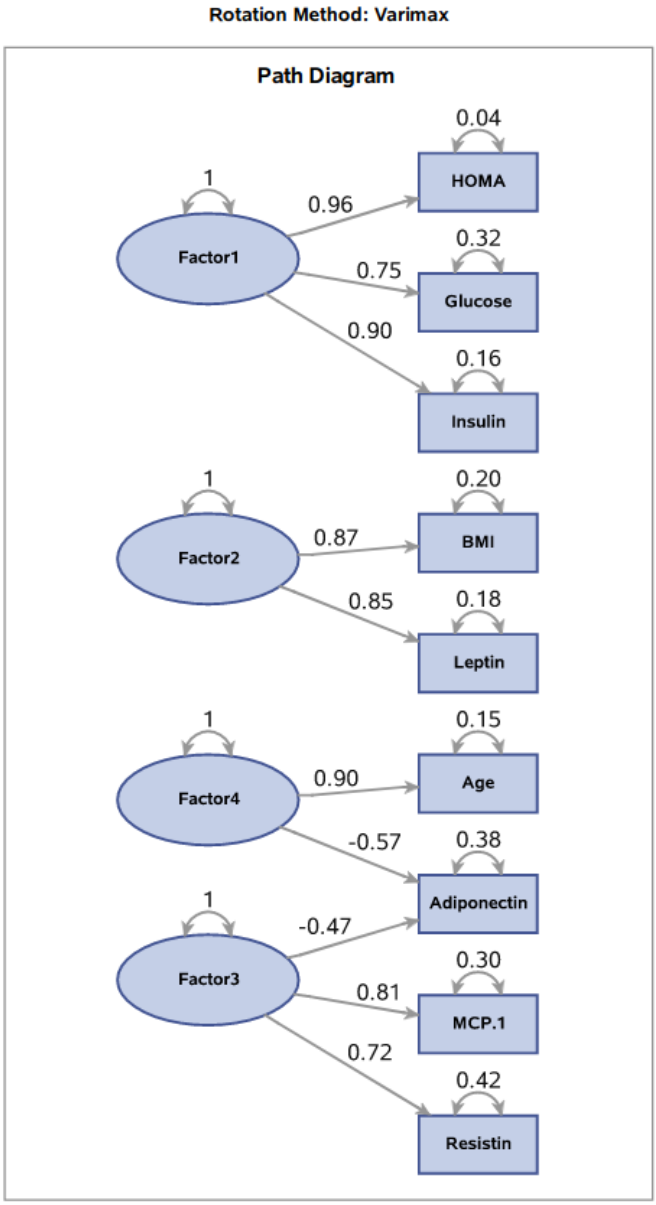

Figure 22: Path Diagram of Factors and Variables.

Naming Factors

We concluded the Principal Component Analysis with 4 components extracted from the original variables. The following names were used to reperesent the factors and a sample from the new dataset created from the factor scores can be found here in the appendix.

Factor Pattern interpretation	
Factor	**Factor Name**
Factor 1	Homa Factors (Wallace et al., 2004)
Factor 2	Metabolism factors (Masoud et al., 2006)
Factor 3	Protein Factors (Resistin, 2021),(Lacolley et al., 2015),(Adiponectin 2021)
Factor 4	Age factors (Obata et al., 2013)

Table 5: New Factor Names

Testing for Correlation Among the Components

A correlation test was done on the 4 components extracted using Pearson's Correlation and Scatter Plot Matrix.

Correlation

The outputs from the correlation test in Figure 23 provide enough evidence that performing Principal Component Analysis on the coimbra breast cancer data would reduce the dimensionality of the dataset to smaller noncorrelated variables.

Pearson Correlation Coefficients, N = 116				
	HOMA FACTORS	METABOLISM FACTORS	PROTEIN FACTORS	AGE FACTORS
HOMA FACTORS	1.00000	0.00000	0.00000	0.00000
METABOLISM FACTORS	0.00000	1.00000	0.00000	0.00000
PROTEIN FACTORS	0.00000	0.00000	1.00000	0.00000
AGE FACTORS	0.00000	0.00000	0.00000	1.00000

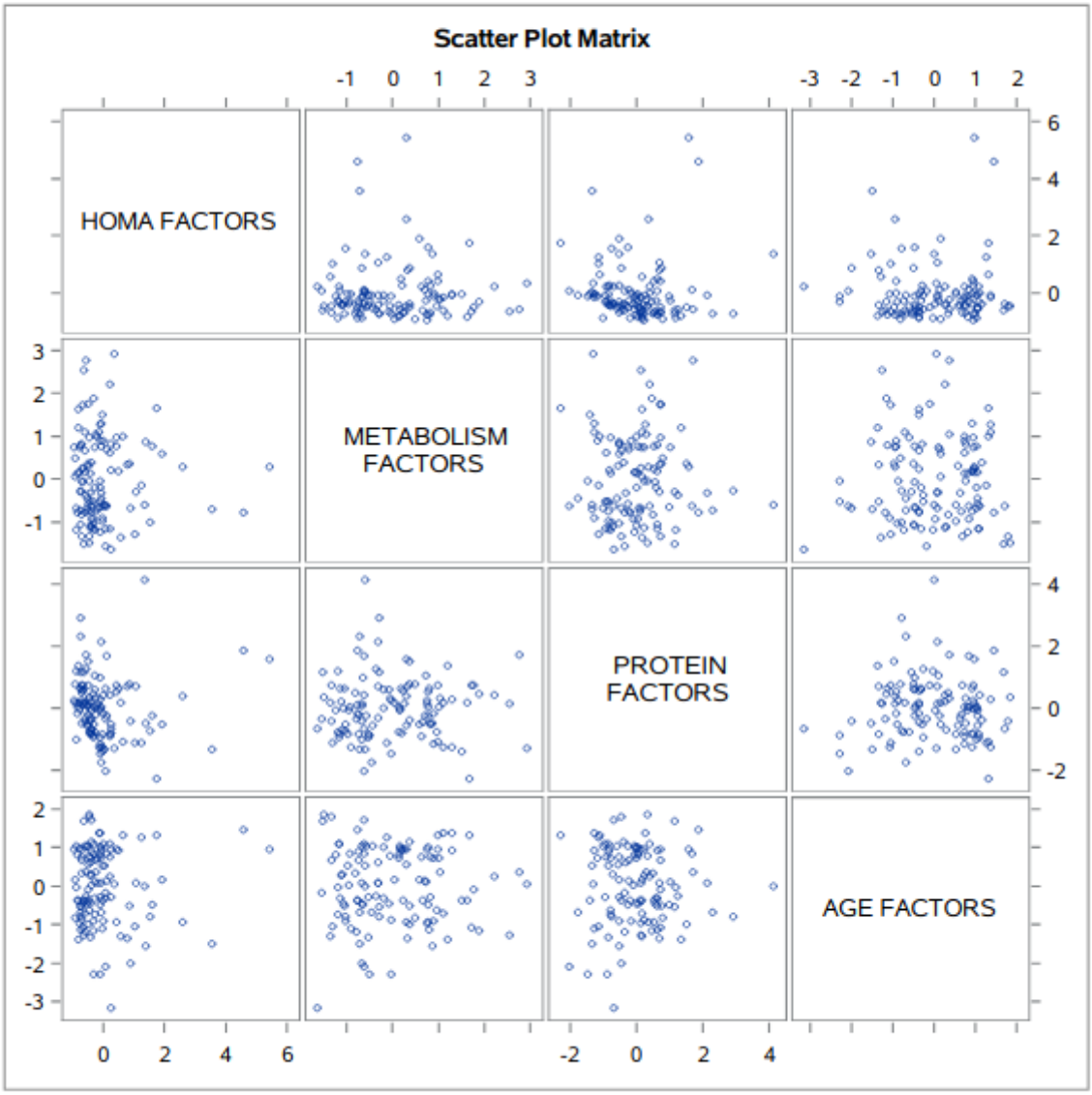

Figure 23: Pearson's Correlation Coefficient and Scatterplot Matrix.

2.3 Research 3: Find the optimal Clustering Technique for classifying observations in the Wisconsin Original Breast Cancer Data into clusters of benign and malignant breast masses between K-Means and Hiearchical clustering.

Dataset Description: The dataset used for performing the cluster analysis in this study was the Breast Cancer Wisconsin original Data from the University of Wisconsin Hospitals Madison, Wisconsin, USA found here. The dataset has 699 observations and 10 variables as seen in Figure 25. In this study we will focus on using cluster analysis to create groups based on the features and class types in this dataset, we will be using 50 random samples from the dataset.

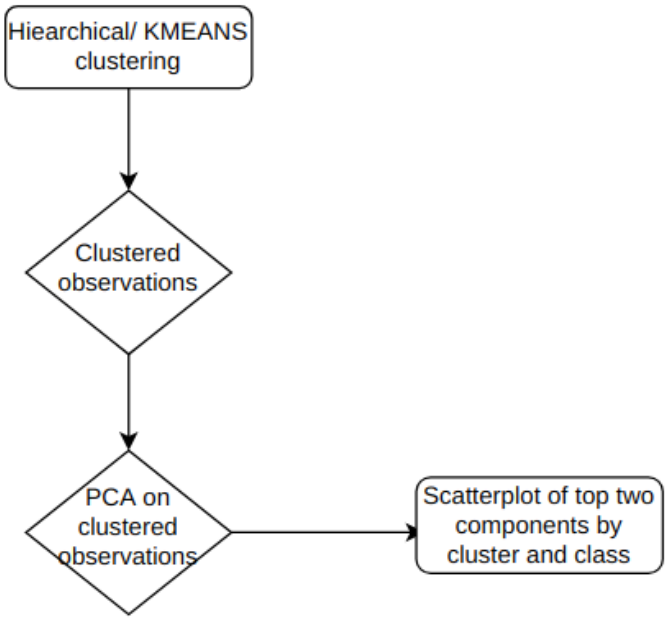

Figure 24: Architecture of Cluster Analysis.

The dataset includes 11 variables as seen in Figure 9, 9 of the variables were used which inludes all variables except Class and Sample code number which are class type and Patient ID respectively.

List of Variables and Attributes	
Variable	**Type**
Sample code number	Num
Clump Thickness	Num.
Uniformity of Cell Size	Num
Uniformity of Cell Shape	Num
Marginal Adhesion	Num
Single Epithelial Cell Size	Num
Bare Nuclei	Num
Bland Chromatin	Num
Normal Nucleoli	Num
Mitoses	Num
Class	Num

Figure 25: Variable Information For Breast Cancer Wisconsin Original Data.

Exploratory Analysis

A summary of the dataset shows that we have 11 variables in the dataset and 50 random instances from the original datasets. The clustering algorithm performs better if we have uncorrelated variables, using the scatterplot matrix in Figure 26 we observe no correlation amongst the variables in the dataset and we will proceed with the clustering of the dataset.

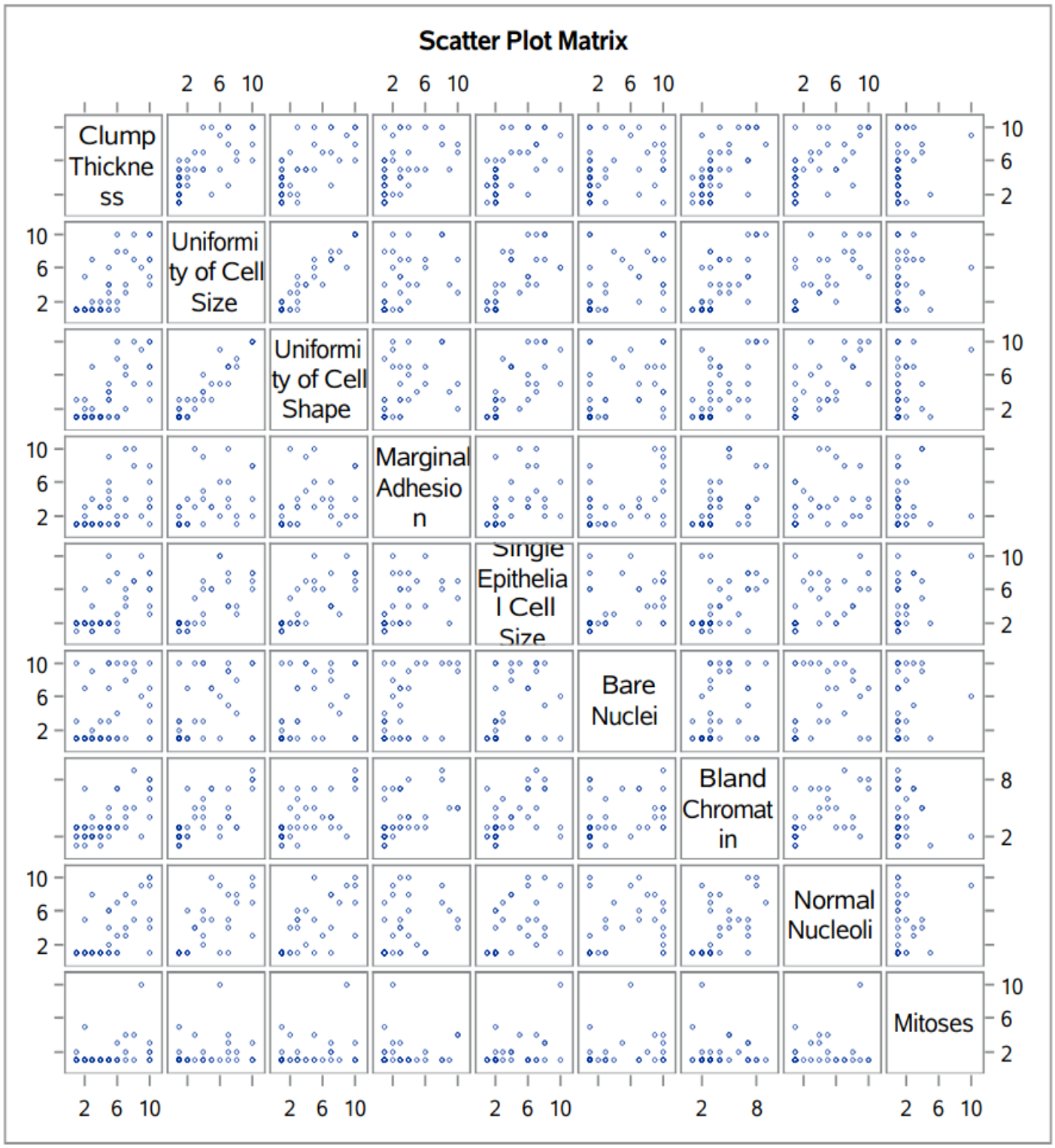

Figure 26: Scatter Plot of the Variables in the Dataset.

Cluster Analysis using Hierarchical Clustering and Standardizing with the Mean

Cluster History

The cluster history table found here in appendix shows the important statistics required for selecting the optimum number of clusters.

Statistics for Variables Interpretation	
Statistic	**Interpretation**
R-square	R-square is the distance between two clusters, this value should be as close to 1 as possible. It explains the proportion of variance in the dataset accounted for by the cluster.
Pseudo F Statistic	The Pseudo-F-statistic shows the closeness of our clusters which is the ratio of the mean sum of squares between the groups to the mean sum of squares within each group. This means we want a high number for this statistic.
Semipartial r-squared	The Semipartial r-squared measures the loss of uniformity due to the merging of the two groups and we need a value that is small for this.
Pseudo t-squared	The Pseudo t-squared statistics we need to locate the point of large difference between the values.

Table 6: Statistics for variable interpretation

Based on the criteria above we can see from the cluster history table here in appendix that the number of clusters that fits this best is 2 clusters as seen in Figure 27. This decision is also confirmed by the denogram in Figure 28.

25	OB37	OB45	2	14.8492	0.0003	.999	1489	.	
24	OB4	OB50	2	15.5563	0.0003	.999	1118	.	
23	OB13	OB26	2	16.2635	0.0003	.999	907	.	
22	CL27	OB9	14	6.0406	0.0005	.998	737	35.1	
21	OB15	OB21	2	22.6274	0.0007	.998	587	.	T
20	OB16	OB44	2	24.0416	0.0007	.997	490	.	
19	CL26	CL22	27	7.5325	0.0010	.996	405	29.6	
18	CL24	OB47	3	23.6784	0.0011	.995	349	3.6	
17	CL21	OB39	3	29.3939	0.0016	.993	295	2.4	
16	OB2	OB40	2	35.3553	0.0016	.991	263	.	T
15	OB38	OB42	2	35.3553	0.0016	.990	244	.	
14	OB22	OB33	2	39.5980	0.0020	.988	225	.	
13	OB6	OB43	2	45.9619	0.0027	.985	204	.	
12	CL23	CL15	4	39.0587	0.0039	.981	180	4.0	
11	CL17	OB19	4	42.4843	0.0047	.976	161	4.3	
10	CL16	OB7	3	50.0000	0.0048	.972	152	3.0	
9	CL18	CL14	5	47.2832	0.0081	.964	135	7.0	
8	CL9	CL13	7	60.6881	0.0142	.949	112	5.0	
7	CL12	CL20	6	60.6424	0.0170	.932	98.7	10.3	
6	CL8	OB24	8	76.2724	0.0240	.908	87.2	5.1	
5	CL10	CL7	9	72.8837	0.0246	.884	85.5	5.7	
4	CL6	CL25	10	83.4210	0.0279	.856	91.0	4.2	
3	CL19	CL5	36	47.3622	0.0445	.811	101	26.7	
2	CL4	CL11	14	96.0871	0.0669	.744	140	9.2	
1	CL3	CL2	50	125.9	0.7444	.000	.	140	

Figure 27: Excerpt from Cluster History.

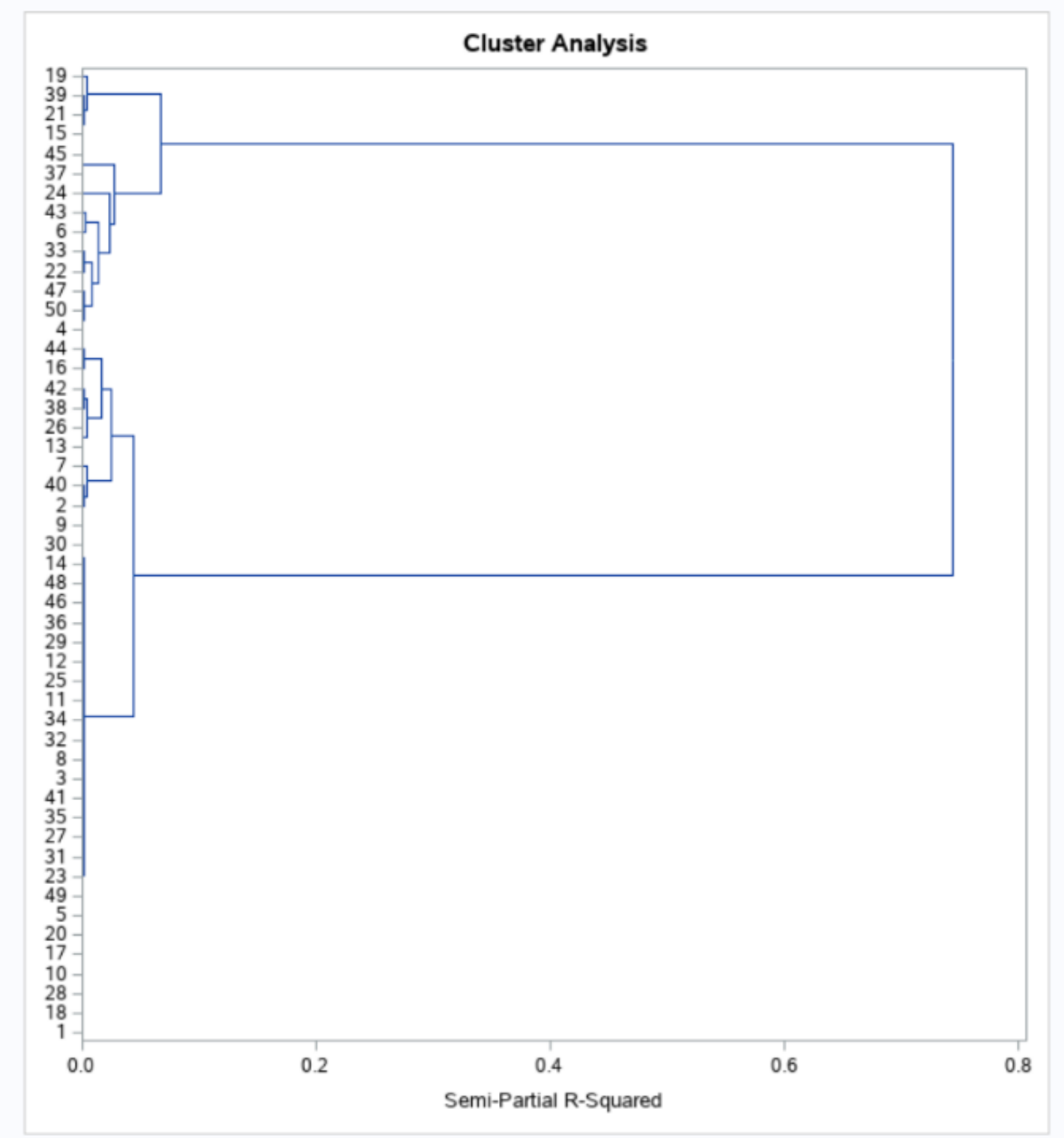

Figure 28: Denogram Clustering the Dataset

New Dataset from Clustering

A new dataset was created using the number of clusters—2 which were derived from the analysis above. The new dataset includes the individual observations and their respective clusters which can be found here in the appendix. Principal component analysis was performed on the data to create scatter plots comparing the first two components grouped by the clusters and class type (benign and malignant breast masses) respectively.

Scatter Plots

The scatterplots in Figure 29 and 30 show that although the hiearchical clustering was able to cluster the observations into clusters of benign and malignant breast mass to $\approx 96\%$ as seen from Figure 29, we still have 4% of the data wrongly grouped as observed by comparing Figure 29 and 30.

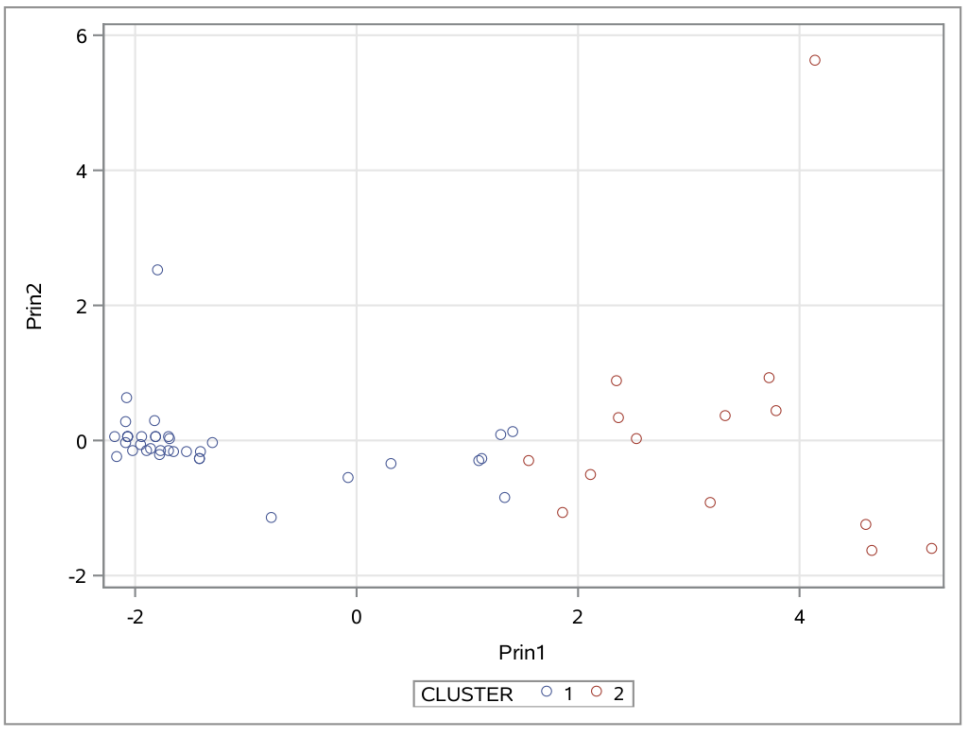

Figure 29: Scatterplot of Principal Componet 1 and 2 Grouped by Cluster.

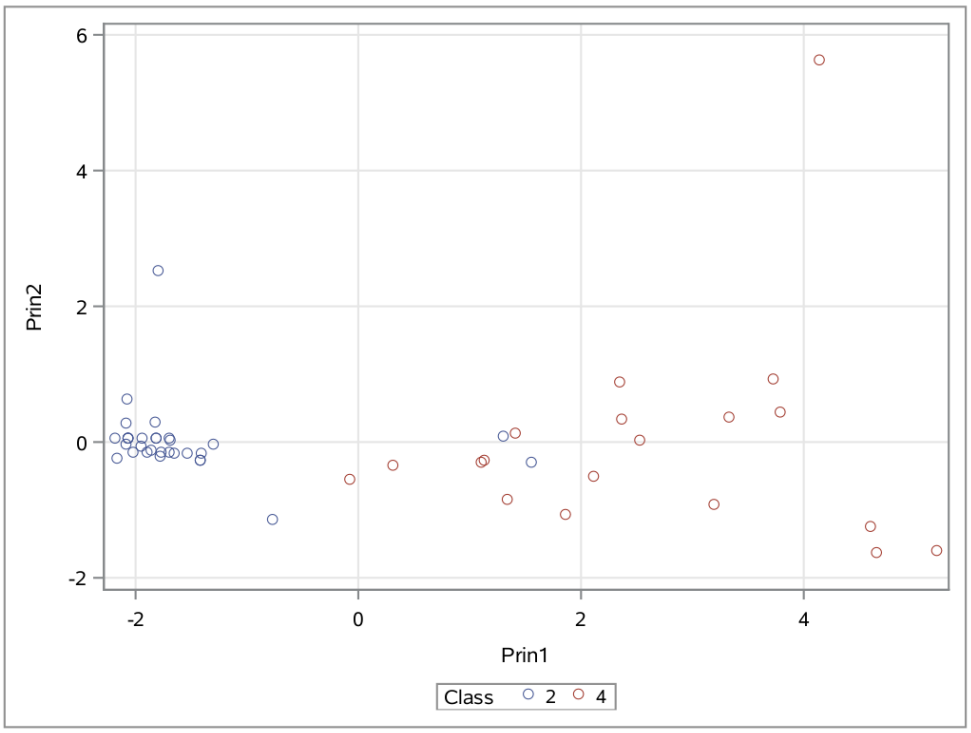

Figure 30: Scatterplot of Principal Componet 1 and 2 Grouped by Breast Masses Type.

Cluster Analysis using K-Means Clustering

The K-Means algorithm was used to cluster the dataset into two clusters, We used two clusters in the K-Means based on the outcome of the hierarchical clustering.

Initial Seeds

The initial seeds provides information about how the cluster was generated at random. The intial seed sets the centroid of each cluster.

Replace=FULL Radius=0 Maxclusters=2 Maxiter=1

Initial Seeds									
Cluster	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	0.111111111	0.000000000	0.000000000	0.000000000	0.111111111	0.000000000	0.000000000	0.000000000	0.444444444
2	0.777777778	1.000000000	1.000000000	0.777777778	0.666666667	1.000000000	1.000000000	0.666666667	0.000000000

Criterion Based on Final Seeds =	0.2054

Figure 31: Table of Initial Seeds by the K-Means Algorithm

Cluster Summary

Figure 32 provides summarized information about the clusters, The distance between the centroid of the nearest cluster and the centroid of the current cluster is the same for both clusters since we have just two clusters.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	32	0.1327	0.9504		2	1.3875
2	18	0.2982	1.1910		1	1.3875

Figure 32: Table of the Cluster Summary

Statistics for Variables

The statistics for variables output displays the statistics of the contributing variables to the clusters. The R-Square and RSQ/(1-RSQ) are two important statistics we focused on from this output.

Statistics for Variables interpretation	
Statistic	**Interpretation**
R-square	The R-Square as defined earlier is the meausre of the difference between the two clusers and we need this value to be close to 1 for each of the variables since it explains the proportion of variance contributed by the variaable to the cluster. The output in Figure 33 shows the R-Square for each value and we can see some of the variables do not have values close to 1.
RSQ/(1-RSQ)	This statistic gives the ratio of between-clusters variance to within-cluster. The value for this has to be relatively high because we need the distance between our clusters to be higher than the distance between datapoints in a cluster, so similar observations can be grouped into same clusters.
Cubic Clustering Criterion	This statistic compares the deviation of our clusters from the expected distribution. Large positive values of Cubic Clustering Criterion greater than 2 are good for our analysis, because it shows we have a large deviation from a unifrom distribution or zero clusters. Figure 33 shows we have a value of 25.429 which satisfy the requirement for a good clustering solution.

Table 7: KMeans Statistics for Variable Interpretation

Statistics for Variables				
Variable	**Total STD**	**Within STD**	**R-Square**	**RSQ/(1-RSQ)**
Clump Thickness	0.31983	0.22925	0.496724	0.986983
Uniformity of Cell Size	0.33489	0.16361	0.766176	3.276716
Uniformity of Cell Shape	0.33619	0.20041	0.651872	1.872508
Marginal Adhesion	0.28373	0.21767	0.423452	0.734460
Single Epithelial Cell Size	0.27416	0.17123	0.617860	1.616842
Bare Nuclei	0.39091	0.28441	0.481474	0.928543
Bland Chromatin	0.25127	0.19222	0.426768	0.744493
Normal Nucleoli	0.32128	0.21516	0.560667	1.276176
Mitoses	0.16830	0.15984	0.116440	0.131785
OVER-ALL	0.30379	0.20706	0.544931	1.197468

Pseudo F Statistic =	57.48

Approximate Expected Over-All R-Squared =	0.16819

Cubic Clustering Criterion =	25.429

Figure 33: Table of the Statistics of the Variables

New Dataset from K-Means Clustering

A new dataset was created from the output of the K-Means clustering. We performed a principal component analysis on the original variables in the dataset and plotted the first two components on a scatter plot while grouping the datapoints by the cluster created from the K-Means and the breast masses type(benign and malignant).

Scatter Plots

Figure 34 and 35 gave similar outcome to the observed outome for the hierarchical clustering. K-Means clustering created two clusters that grouped the observations into clusters of benign and malignant breast mass as seen in the cluster table here in the appendix to $\approx 96\%$ accuracy as seen from Figure 29.

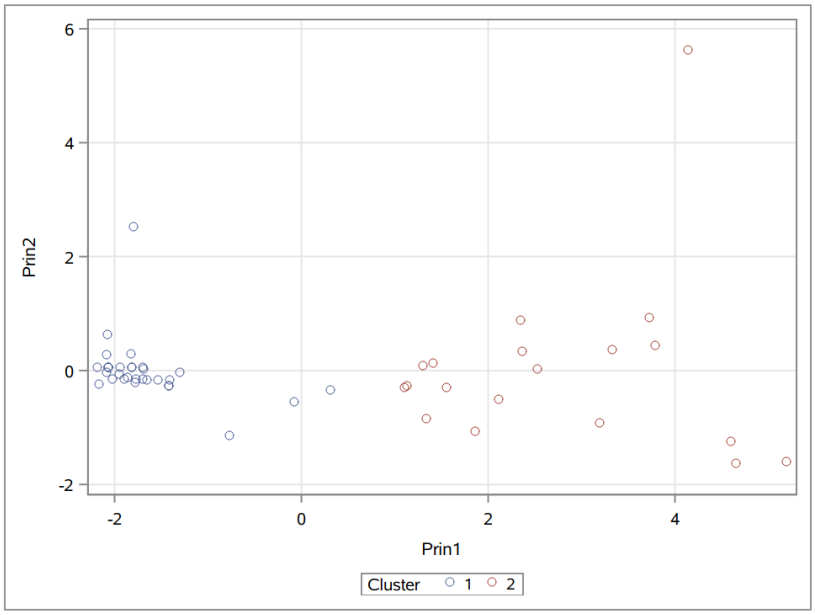

Figure 34: K-means Scatterplot of Principal Componet 1 and 2 Grouped by Cluster

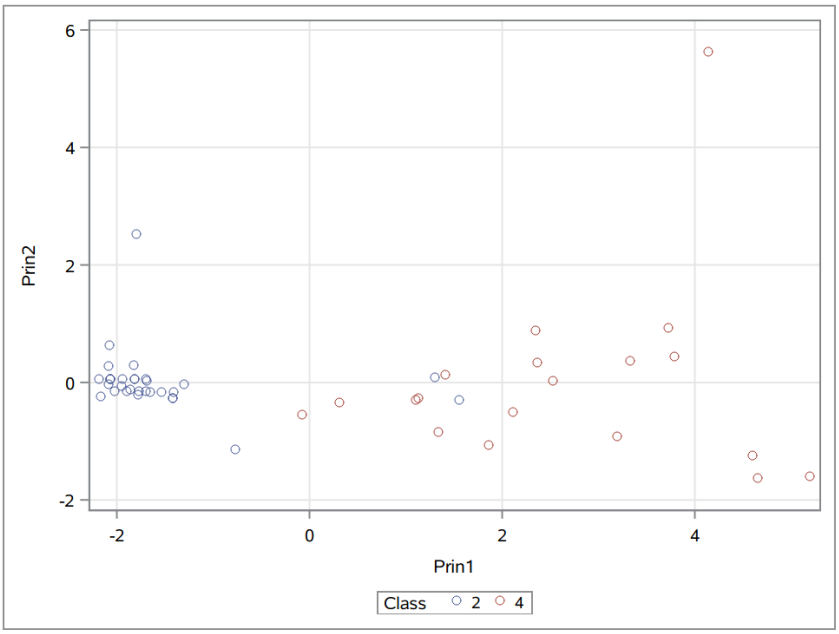

Figure 35: K-Means Scatterplot of Principal Componet 1 and 2 Grouped by Breast Masses Type

3 Results and Discussion

Logistic Regression

We used Logistic Regression to model the relationship between Age and Number of positive axillary nodes detected to model the odds of a patient surviving more than 5 years after the operation. Based on the Model fit statistics using AIC and Schwarz criteria we observed that the value for the Intercept and Covariates is lesser than Intercept only as seen in Figure 3 and for this two test the model with the smallest value is favoured which means our model favours the former. The global null hypothesis test using Likelihood ratio, Score and Wald tests show that our model is better than an empty model which was confirmed by the outcome of the 3 tests At a significance value of 0.05. The analysis of maximum likelihood estimates gave us the equation for our model which is given by :

$$Logit(P) = 2.4629 - 0.0197 * Age - 0.0883 * \text{number of positive axillary nodes detected}$$

The odds ratio estimate was used to understand the changes in the odds of surviving more than 5 years for a unit change in our input variables with a 95% confidence interval. Finally, the Plot for predicted probabilities for Survival shows us the probability of surviving more than 5 years has a downward trend as age increases as seen in Figure 7.

Principal Component Analysis

Principal Component Analysis was used to reduce the dimensionality of the anthropometric dataset. We set a proportion of variance needed for our analysis, this was validated using the Kaiser's rule and the Scree Plots. The varimax Orthogonal rotation was used to maximise the variance of highly correlated loadings and reduce low correlations. The new Factor pattern obtained after rotation showed that we have a more balanced distribution of variance as seen in figure 18. Finally, the components were renamed using succinct and clear words that explains the intrinsic elements of each component extracted and a correlation test confirmed we have no correlation between the components.

Cluster Analysis

The optimum number of clusters was decided using 4 statistics and we discovered the number of clusters that fits the 4 statistics best was 2 clusters as seen in Figure 27 and 28. A new dataset was created with this clusters from both K-Means and Hierarchical clustering and we grouped the observations into their respective clusters, a comparison was done between the accuracy of the K-Means and Hierarchical clustering by using a scatter plot of the principal components of the variables grouping them by clusters and the original class variable in the dataset. We observed that the algorithm was able to classify our observations to some level of accuracy for both K-Means and Hierarchical clustering. K-Means clustering has Approximately, 92% prediction accuracy while hierarchical clustering has approximately 86% seen in here in the appendix.

4 Conclusion

Logistic Regression

- The experimental results revealed that the variables age and Number of positive axillary nodes detected would improve the model for predicting the odds of the patients long term survival.
- We successfully created a model that explains the relationship between the variables, our model showed that age and number of positive axillary nodes detected contributed to the long term survival of cancer patients that had undergone surgery. which is expressed as: $Logit(P) = 2.4629 - 0.0197 * Age - 0.0883 * \text{number of positive axillary nodes}$
- Based on the outcome of this study we confirmed that a change in the variables age and Number of positive axillary nodes detected will impact the odds of survival of the patient and this was represented by the odds ratio estimate table in Figure 5.
- Finally, the study revealed that the probability of surviving more than 5 years after the operation has an inverse relationship with the age of the patient.

Principal Component Analysis

- This study confirmed the ability of PCA to successfully reduce the dimensionality of the Coimbra breast cancer dataset into four noncorrelated components that can be used for further study in the field of cancer research.
- We successfully extracted the principal components of the dataset by applying Kaiser's rule and the Scree plot, our dataset passed these component extraction methods.
- Using varimax Rotation we successfully maximized the highly correlated variables and we renamed the new components using succinct and sensible placeholders which explain the intricate features of each component.

Cluster Analysis

- The process of clustering the Wisconsin dataset was done using Hierarchical and K-Means clustering. This study successfully compared the output of the K-Means and Hierarchical clustering and discovered that K-Means algorithm was more accurate than the Hierarchical clustering. K-Means had a success prediction rate of $\approx 92\%$ while Hierarchical clustering was less accurate with $\approx 86\%$ accuracy. This means to get the best clustering results from this dataset, K-Means clustering is a better option than Hierarchical clustering.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394-424. doi: 10.3322/caac.21492. Epub 2018 Sep 12. Erratum in: CA Cancer J Clin. 2020 Jul;70(4):313. PMID: 30207593.

[2] Cancer Research UK, https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared, (Accessed 29 Dec 2021)

[3] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research, Volume 70, 2017, Pages 263 − 286, ISSN 0148 − 2963, https://doi.org/10.1016/j.jbusres.2016.08.001 (http://www.sciencedirect.com/science/article/pii/S014829631630488X).

[4] Wallace, T., Levy, J. and Matthews, D., 2004. Use And Abuse Of HOMA Modeling.

[5] Wilcox G. Insulin and insulin resistance. Clin Biochem Rev. 2005 May;26(2):19-39. PMID: 16278749; PMCID: PMC1204764.

[6] Masoud Y Al Maskari1, and Adel A Alnaqdy (2006). Correlation between Serum Leptin Levels, Body Mass Index and Obesity in Omanis

[7] Wikipedia contributors. (2021, January 02). Resistin. In Wikipedia, The Free Encyclopedia. Retrieved 11:06, January 02, 2021, from https://en.wikipedia.org/wiki/Resistin

[8] Patrick Lacolley, Pascal Challande, Veronique Regnault, Edward G.Lakatta, Mingyi Wang. Early Vascular Aging (EVA) New Directions in Cardiovascular Protection (2015)

[9] Wikipedia contributors. (2021, January 02). Adiponectin. In Wikipedia, The Free Encyclopedia. Retrieved 11:16, January 02, 2021, from https://en.wikipedia.org/wiki/Adiponectin

[10] Obata Y, Yamada Y, Takahi Y, Baden MY, Saisho K, Tamba S, Yamamoto K, Umeda M, Furubayashi A, Matsuzawa Y. Relationship between serum adiponectin levels and age in healthy subjects and patients with type 2 diabetes, 2013.

APPENDIX

Principal Component Analysis

	HOMA FACTORS	METABOLISM FACTORS	PROTEIN FACTORS	AGE FACTORS
1	-0.85329425	-0.766336886	-0.181706381	-0.358128528
2	-0.434106576	-1.324576701	-0.444304336	1.7908024734
3	-0.039211349	-0.985631317	-0.854276648	0.528480734
4	-0.689082201	-1.322375715	0.7303012875	0.6728908521
5	-0.449987431	-1.49201063	0.3380428467	1.8520989516
6	-0.422823502	-1.068127466	-0.013393457	-0.465598103
7	-0.619040543	-1.516424477	1.1493033025	1.6843210362
8	0.2321117802	-1.15688582	-0.863882323	1.0862472916
9	-0.312820172	-1.156020709	-0.942543525	1.0822813561
10	-0.376389357	-0.738976549	-0.839733278	0.9073069282
11	-0.595678533	-0.787155944	-0.376119387	-1.186829566
12	-0.095987611	-0.038365474	-1.473892546	-2.301250777
13	-0.308506301	-0.513946796	-0.884057486	-2.289347116
14	0.2427821688	-1.639437817	-0.689915047	-3.163463344
15	-0.428300849	-0.528221386	-0.859534579	-1.339772421
16	-0.018537615	-0.908681509	-1.262640033	-1.094777965
17	0.0720682642	-0.631535791	-2.036426716	-2.078706181
18	0.1309292822	0.6933848639	-0.241522549	0.128777963
19	-0.680242179	0.8115456422	-0.512779675	0.7641707291
20	-0.313658176	1.8794464463	0.4558465369	-1.166667716
21	-0.707202371	-0.038951145	0.2367705136	-0.881129824
22	-0.751339454	0.7459833809	0.1163022201	-0.953041673
23	-0.78887698	1.1900505934	1.3451123645	-1.372438961
24	-0.736227825	0.7903670263	0.6178704692	-1.122418008
25	-0.563841079	-0.366194711	1.2468993576	-0.265213241
26	-0.812538614	1.6312126473	0.1585433425	-0.364508315
27	-0.49432929	1.7602623361	0.7243931332	-0.097534286
28	-0.257523856	0.7376066739	1.0241663491	0.1125021167
29	-0.652579925	1.7269626691	0.7250030234	-1.063456655
30	-0.90358093	0.486866699	0.7769820892	-0.81527134

Figure 36: 30 Sample Observations from Principal Components of Coimbra Dataset

Root-Mean-Square Distance Between Observations	178.061

Cluster History

Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
49	OB12	OB29	2	0	0.0000	1.00	.	.	T
48	CL49	OB36	3	0	0.0000	1.00	.	.	T
47	OB5	OB49	2	0	0.0000	1.00	.	.	
46	OB10	OB17	2	0.7071	0.0000	1.00	14E4	.	T
45	OB1	OB18	2	0.7071	0.0000	1.00	88E3	.	T
44	OB11	OB25	2	0.7071	0.0000	1.00	72E3	.	T
43	OB23	OB31	2	0.7071	0.0000	1.00	65E3	.	T
42	OB8	OB32	2	0.7071	0.0000	1.00	61E3	.	T
41	OB46	OB48	2	0.7071	0.0000	1.00	58E3	.	
40	CL43	OB27	3	1.0000	0.0000	1.00	44E3	3.0	T
39	CL45	OB28	3	1.0000	0.0000	1.00	37E3	3.0	T
38	CL42	OB34	3	1.0000	0.0000	1.00	34E3	3.0	
37	CL48	CL41	5	0.8944	0.0000	1.00	28E3	16.2	
36	CL40	OB35	4	1.2910	0.0000	1.00	24E3	3.0	
35	CL39	CL46	5	1.2247	0.0000	1.00	21E3	4.2	
34	OB3	CL38	4	1.5275	0.0000	1.00	17E3	5.0	
33	CL44	CL37	7	1.3274	0.0000	1.00	14E3	9.3	
32	CL35	OB20	6	1.9322	0.0000	1.00	11E3	8.4	
31	CL34	CL33	11	1.8974	0.0000	1.00	8245	9.4	
30	CL31	OB14	12	2.3484	0.0000	1.00	6352	6.9	
29	CL36	OB41	5	3.3015	0.0000	1.00	4738	23.2	
28	CL32	CL47	8	3.0355	0.0001	1.00	3750	14.7	
27	CL30	OB30	13	3.1744	0.0001	1.00	3000	10.9	
26	CL28	CL29	13	4.0982	0.0001	1.00	2312	9.5	
25	OB37	OB45	2	14.8492	0.0003	.999	1489	.	
24	OB4	OB50	2	15.5563	0.0003	.999	1118	.	
23	OB13	OB26	2	16.2635	0.0003	.999	907	.	
22	CL27	OB9	14	6.0406	0.0005	.998	737	35.1	
21	OB15	OB21	2	22.6274	0.0007	.998	587	.	T
20	OB16	OB44	2	24.0416	0.0007	.997	490	.	
19	CL26	CL22	27	7.5325	0.0010	.996	405	29.6	
18	CL24	OB47	3	23.6784	0.0011	.995	349	3.6	
17	CL21	OB39	3	29.3939	0.0016	.993	295	2.4	
16	OB2	OB40	2	35.3553	0.0016	.991	263	.	T
15	OB38	OB42	2	35.3553	0.0016	.990	244	.	
14	OB22	OB33	2	39.5980	0.0020	.988	225	.	
13	OB6	OB43	2	45.9619	0.0027	.985	204	.	
12	CL23	CL15	4	39.0587	0.0039	.981	180	4.0	
11	CL17	OB19	4	42.4843	0.0047	.976	161	4.3	
10	CL16	OB7	3	50.0000	0.0048	.972	152	3.0	
9	CL18	CL14	5	47.2832	0.0081	.964	135	7.0	
8	CL9	CL13	7	60.6881	0.0142	.949	112	5.0	
7	CL12	CL20	6	60.6424	0.0170	.932	98.7	10.3	
6	CL8	OB24	8	76.2724	0.0240	.908	87.2	5.1	
5	CL10	CL7	9	72.8837	0.0246	.884	85.5	5.7	
4	CL6	CL25	10	83.4210	0.0279	.856	91.0	4.2	
3	CL19	CL5	36	47.3622	0.0445	.811	101	26.7	
2	CL4	CL11	14	96.0871	0.0669	.744	140	9.2	
1	CL3	CL2	50	125.9	0.7444	.000	.	140	

Figure 37: Cluster History Table for Hierarchical Clustering

CLUSTER=1

Obs	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1	1036172	2	1	1	1	2	1	2	1	1	2
2	1067444	2	1	1	1	2	1	2	1	1	2
3	1079304	2	1	1	1	2	1	2	1	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2
5	1106095	4	1	1	3	2	1	3	1	1	2
6	1033078	4	2	1	1	2	1	2	1	1	2
7	1048672	4	1	1	1	2	1	2	1	1	2
8	1000025	5	1	1	1	2	1	3	1	1	2
9	1049815	4	1	1	1	2	1	3	1	1	2
10	1035283	1	1	1	1	1	1	3	1	1	2
11	1059552	1	1	1	1	2	1	3	1	1	2
12	1056784	3	1	1	1	2	1	2	1	1	2
13	1070935	3	1	1	1	1	1	2	1	1	2
14	1018561	2	1	2	1	2	1	3	1	1	2
15	1071760	2	1	1	1	2	1	3	1	1	2
16	1103722	1	1	1	1	2	1	2	1	2	2
17	1105524	1	1	1	1	2	1	2	1	1	2
18	1066373	3	2	1	1	1	1	2	1	1	2
19	1066979	5	1	1	1	2	1	2	1	1	2
20	1074610	2	1	1	2	2	1	3	1	1	2
21	1075123	3	1	2	1	2	1	2	1	1	2
22	1015425	3	1	1	1	2	2	3	1	1	2
23	1050718	6	1	1	1	2	1	3	1	1	2
24	1043999	1	1	1	1	2	3	3	1	1	2
25	1190394	4	1	1	1	2	3	1	1	1	2
26	1070935	1	1	3	1	2	1	1	1	1	2
27	1041801	5	3	3	3	2	3	4	4	1	4
28	1065726	5	2	3	4	2	7	3	6	1	4
29	1033078	2	1	1	1	2	1	1	1	5	2
30	1047630	7	4	6	4	6	1	4	3	1	4
31	1102573	5	6	5	6	10	1	3	1	1	4
32	1002945	5	4	4	5	7	10	3	2	1	2
33	1091262	2	5	3	3	6	7	7	5	1	4
34	1081791	6	2	1	1	1	1	7	1	1	2
35	1099510	10	4	3	1	3	3	6	5	2	4
36	1018099	1	1	1	1	2	10	3	1	1	2

Figure 38: Hierarchical Clustering Data for Cluster 1

CLUSTER=2

Obs	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
37	1080185	10	10	10	8	6	1	8	9	1	4
38	1103608	10	10	10	4	8	1	8	10	1	4
39	1016277	6	8	8	1	3	4	3	7	1	2
40	1106829	7	8	7	2	4	8	3	8	2	4
41	1044572	8	7	5	10	7	9	5	5	4	4
42	1054590	7	3	2	10	5	10	5	4	4	4
43	1105257	3	7	7	4	4	9	4	8	1	4
44	1084584	5	4	4	9	2	10	5	6	1	4
45	1054593	10	5	5	3	6	7	7	10	1	4
46	1072179	10	7	7	3	8	5	7	4	3	4
47	1017122	8	10	10	8	7	10	9	7	1	4
48	1100524	6	10	10	2	8	10	7	3	3	4
49	1050670	10	7	7	6	4	10	4	1	2	4
50	1165926	9	6	9	2	10	6	2	9	10	4

Figure 39: Hierarchical Clustering Data for Cluster 2

Cluster=1

Obs	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1	1000025	0.1389616668	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
2	1015425	-0.555846667	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.42635749	-0.298481003	-0.664013467	-0.356508599	2
3	1017023	-0.2084425	-0.743203086	-0.733721189	0.1409780558	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
4	1018099	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.583605461	1.8475491218	-0.298481003	-0.664013467	-0.356508599	2
5	1018561	-0.903250834	-0.743203086	-0.403216149	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
6	1033078	-0.903250834	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-1.293417679	-0.664013467	2.2842958372	2
7	1033078	-0.2084425	-0.411415994	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
8	1035283	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.988887031	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
9	1036172	-0.903250834	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
10	1041801	0.1389616668	-0.079628902	-0.072711109	0.1409780558	-0.583605461	-0.142119163	0.1989873353	0.373507575	-0.356508599	4
11	1043999	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.142119163	-0.298481003	-0.664013467	-0.356508599	2
12	1048672	-0.2084425	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
13	1049815	-0.2084425	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
14	1050718	0.4863658336	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
15	1056784	-0.555846667	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
16	1059552	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
17	1065726	0.1389616668	-0.411415994	-0.072711109	0.5325837664	-0.583605461	0.9948341425	-0.298481003	1.0651882696	-0.356508599	4
18	1066373	-0.555846667	-0.411415994	-0.733721189	-0.642233365	-0.988887031	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
19	1066979	0.1389616668	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
20	1067444	-0.903250834	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
21	1070935	-1.250655001	-0.743203086	-0.072711109	-0.642233365	-0.583605461	-0.710595816	-1.293417679	-0.664013467	-0.356508599	2
22	1070935	-0.555846667	-0.743203086	-0.733721189	-0.642233365	-0.988887031	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
23	1071760	-0.903250834	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
24	1074610	-0.903250834	-0.743203086	-0.733721189	-0.250627655	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2
25	1075123	-0.555846667	-0.743203086	-0.403216149	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
26	1079304	-0.903250834	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
27	1081791	0.4863658336	-0.411415994	-0.733721189	-0.642233365	-0.988887031	-0.710595816	1.6913923498	-0.664013467	-0.356508599	2
28	1190394	-0.2084425	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.142119163	-1.293417679	-0.664013467	-0.356508599	2
29	1103722	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	0.3036925102	2
30	1105524	-1.250655001	-0.743203086	-0.733721189	-0.642233365	-0.583605461	-0.710595816	-0.795949341	-0.664013467	-0.356508599	2
31	1106095	-0.2084425	-0.743203086	-0.733721189	0.1409780558	-0.583605461	-0.710595816	-0.298481003	-0.664013467	-0.356508599	2

Figure 40: K-Means Clustering Data for Cluster 1

Cluster=2

Obs	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
32	1002945	0.1389616668	0.2521581898	0.2577939313	0.9241894769	1.4428023892	1.8475491218	-0.298481003	-0.318173119	-0.356508599	2
33	1016277	0.4863658336	1.5793065573	1.5798140916	-0.642233365	-0.178323891	0.1421191632	-0.298481003	1.4110286168	-0.356508599	2
34	1017122	1.1811741674	2.2428807411	2.2408241718	2.0990066086	1.4428023892	1.8475491218	2.6863290261	1.4110286168	-0.356508599	4
35	1044572	1.1811741674	1.2475194655	0.5882989714	2.8822180298	1.4428023892	1.5633107954	0.6964556734	0.7193479223	1.6240947282	4
36	1047630	0.8337700005	0.2521581898	0.9188040114	0.5325837664	1.0375208192	-0.710595816	0.1989873353	0.0276672278	-0.356508599	4
37	1050670	1.8759825012	1.2475194655	1.2493090515	1.3157951875	0.2269576792	1.8475491218	0.1989873353	-0.664013467	0.3036925102	4
38	1054590	0.8337700005	-0.079628902	-0.403216149	2.8822180298	0.6322392492	1.8475491218	0.6964556734	0.373507575	1.6240947282	4
39	1054593	1.8759825012	0.5839452817	0.5882989714	0.1409780558	1.0375208192	0.9948341425	1.6913923498	2.4485496586	-0.356508599	4
40	1165926	1.5285783343	0.9157323736	1.9103191317	-0.250627655	2.6586470992	0.7105958161	-0.795949341	2.1027093113	5.5853013824	4
41	1072179	1.8759825012	1.2475194655	1.2493090515	0.1409780558	1.8480839592	0.4263574897	1.6913923498	0.373507575	0.9638936192	4
42	1080185	1.8759825012	2.2428807411	2.2408241718	2.0990066086	1.0375208192	-0.710595816	2.1888606879	2.1027093113	-0.356508599	4
43	1084584	0.1389616668	0.2521581898	0.2577939313	2.4906123192	-0.583605461	1.8475491218	0.6964556734	1.0651882696	-0.356508599	4
44	1091262	-0.903250834	0.5839452817	-0.072711109	0.1409780558	1.0375208192	0.9948341425	1.6913923498	0.7193479223	-0.356508599	4
45	1099510	1.8759825012	0.2521581898	-0.072711109	-0.642233365	-0.178323891	-0.142119163	1.1939240116	0.7193479223	0.3036925102	4
46	1100524	0.4863658336	2.2428807411	2.2408241718	-0.250627655	1.8480839592	1.8475491218	1.6913923498	0.0276672278	0.9638936192	4
47	1102573	0.1389616668	0.9157323736	0.5882989714	1.3157951875	2.6586470992	-0.710595816	-0.298481003	-0.664013467	-0.356508599	4
48	1103608	1.8759825012	2.2428807411	2.2408241718	0.5325837664	1.8480839592	-0.710595816	2.1888606879	2.4485496586	-0.356508599	4
49	1105257	-0.555846667	1.2475194655	1.2493090515	0.5325837664	0.2269576792	1.5633107954	0.1989873353	1.7568689641	-0.356508599	4
50	1106829	0.8337700005	1.5793065573	1.2493090515	-0.250627655	0.2269576792	1.279072469	-0.298481003	1.7568689641	0.3036925102	4

Figure 41: K-Means Clustering Data for Cluster 2