Breast Cancer Patient Review

ALYSSA WHEELER

Introduction

- Breast Cancer Dataset
- Attributes to look at: age, tumor stage, patient status, histology
- Source: https://www.kaggle.com/datasets/amandam1/breastcancerdataset CSV called “BRCA”. Located on Kaggle and can find in the download tab on the top right.

Variable Dictionary

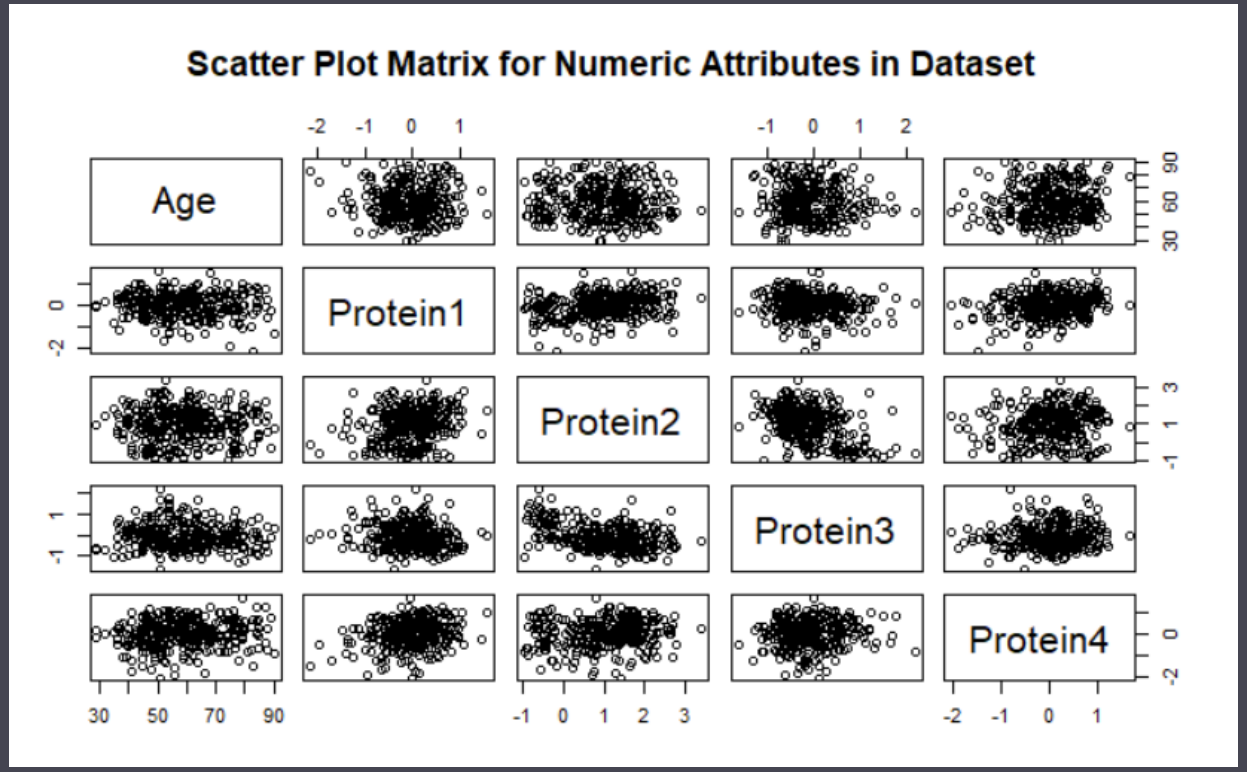
- Patient_ID: unique identifier id of a patient- Arbitrary
- Age (29-90): age at diagnosis (Years)- Continuous
- Gender: Male/Female- Categorical
- Protein1, Protein2, Protein3, Protein4: expression levels (undefined units)- Continuous
- Tumour_Stage: I, II, III - Ranked
- Histology: Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, Mucinous Carcinoma- Categorical
- ER status: Estrogen receptor hormone to test if breast cancer cells are Positive/Negative- Boolean
- PR status: Progesterone receptor hormone to test if breast cancer cells are Positive/Negative- Boolean
- HER2 status: Human epidermal growth factor receptor 2 protein to test if breast cancer cells are Positive/Negative- Boolean
- Surgery_type: Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other- Categorical
- Date_of_Surgery: Date on which surgery was performed (in YY-MM-DD)- Chronological
- Date_of_Last_Visit: Date of last visit (in YY-MM-DD) [null, in case the patient didn’t visited again after the surgery]- Chronological
- Patient_Status: Alive/Dead [null, in case the patient didn’t visited again after the surgery and there is no information available whether the patient is alive or dead]. - Categorical
- DateDiff: difference in days between the date of surgery and the last visit

Data

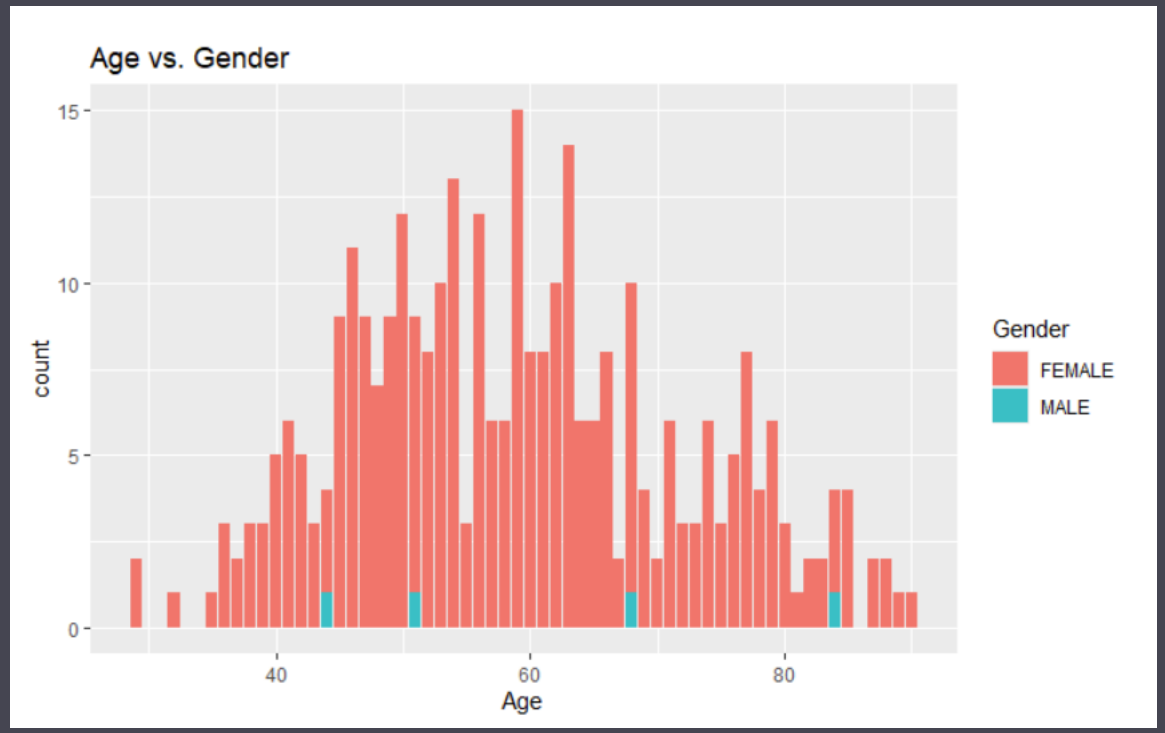
Patient_ID <chr>	Age <dbl>	Gender <chr>	Protein1 <dbl>	Protein2 <dbl>	Protein3 <dbl>	Protein4 <dbl>	Tumour_Stage <chr>	Histology <chr>
TCGA-D8-A1XD	36	FEMALE	0.08035300	0.4263800	0.5471500	0.27368000	III	Infiltrating Ductal Carcinoma
TCGA-EW-A1OX	43	FEMALE	-0.42032000	0.5780700	0.6144700	-0.03150500	II	Mucinous Carcinoma
TCGA-A8-A079	69	FEMALE	0.21398000	1.3114000	-0.3274700	-0.23426000	III	Infiltrating Ductal Carcinoma
TCGA-D8-A1XR	56	FEMALE	0.34509000	-0.2114700	-0.1930400	0.12427000	II	Infiltrating Ductal Carcinoma
TCGA-BH-A0BF	56	FEMALE	0.22155000	1.9068000	0.5204500	-0.31199000	II	Infiltrating Ductal Carcinoma
TCGA-AO-A1KQ	84	MALE	-0.08187200	1.7241000	-0.0573350	0.04302500	III	Infiltrating Ductal Carcinoma
TCGA-D8-A73X	53	FEMALE	-0.06953500	1.4183000	-0.3610500	0.39158000	II	Infiltrating Ductal Carcinoma
TCGA-EW-A1P5	77	FEMALE	-0.15175000	-0.6633200	1.1894000	0.21718000	II	Infiltrating Ductal Carcinoma
TCGA-A8-A09A	40	FEMALE	-0.56570000	1.2668000	-0.2934600	0.19395000	II	Infiltrating Lobular Carcinoma
TCGA-S3-A6ZG	71	FEMALE	-0.22305000	0.5059400	-0.3494300	-0.83530000	II	Infiltrating Ductal Carcinoma

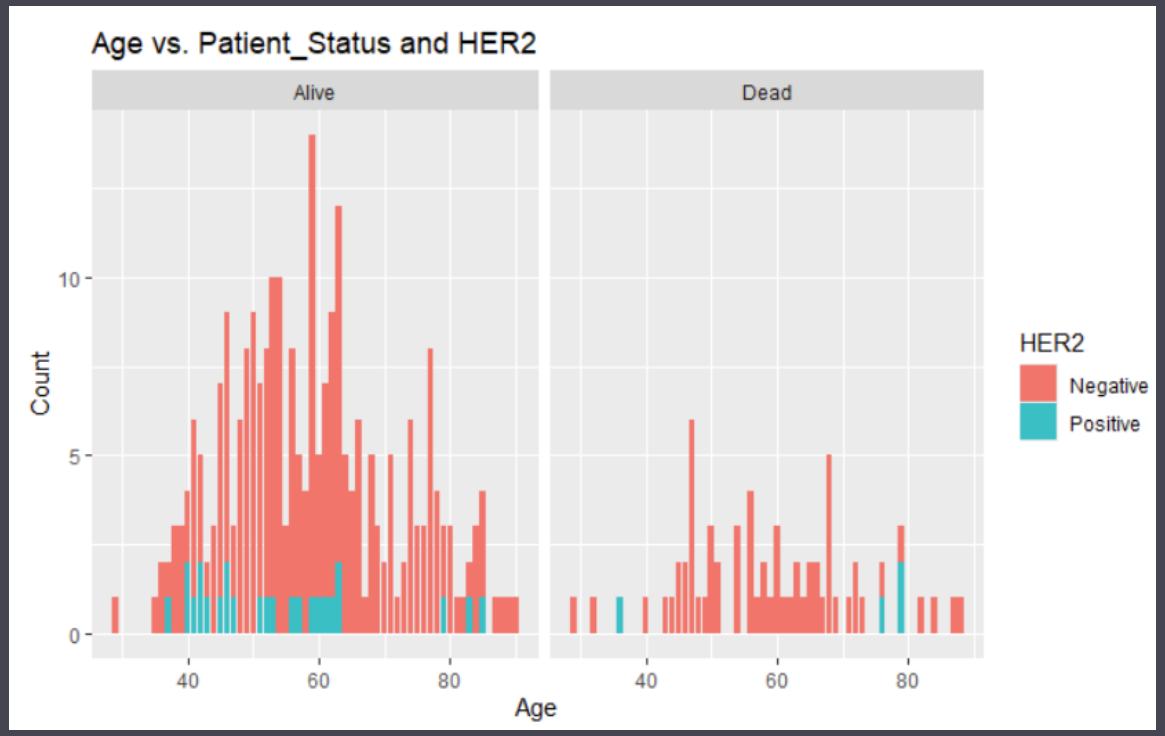
ER <chr>	PR <chr>	HER2 <chr>	Surgery_type <chr>	Date_of_Surgery <chr>	Date_of_Last_Visit <chr>	Patient_Status <chr>	DateDiff <dbl>
Positive	Positive	Negative	Modified Radical Mastectomy	17-01-15	17-06-19	Alive	155
Positive	Positive	Negative	Lumpectomy	17-04-26	18-11-09	Dead	562
Positive	Positive	Negative	Other	17-09-08	18-06-09	Alive	274
Positive	Positive	Negative	Modified Radical Mastectomy	17-01-25	17-07-12	Alive	168
Positive	Positive	Negative	Other	17-05-06	19-06-27	Dead	782
Positive	Positive	Negative	Modified Radical Mastectomy	17-09-18	21-11-15	Alive	1519
Positive	Positive	Negative	Simple Mastectomy	17-02-04	18-02-07	Alive	368
Positive	Positive	Negative	Modified Radical Mastectomy	17-09-28	18-09-28	Alive	365
Positive	Positive	Positive	Other	17-02-14	17-12-15	Alive	304
Positive	Positive	Negative	Lumpectomy	17-05-26	17-12-19	Alive	207

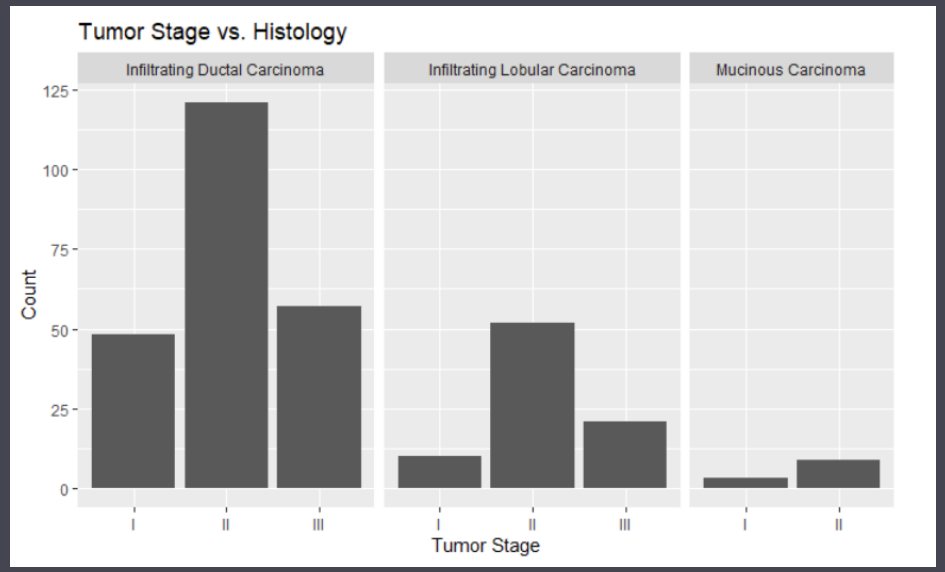
Scatterplot Matrix

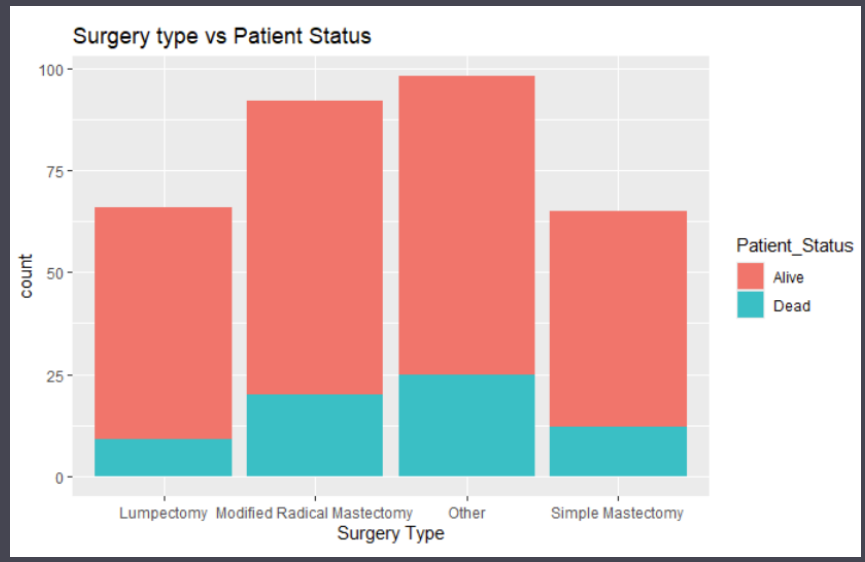


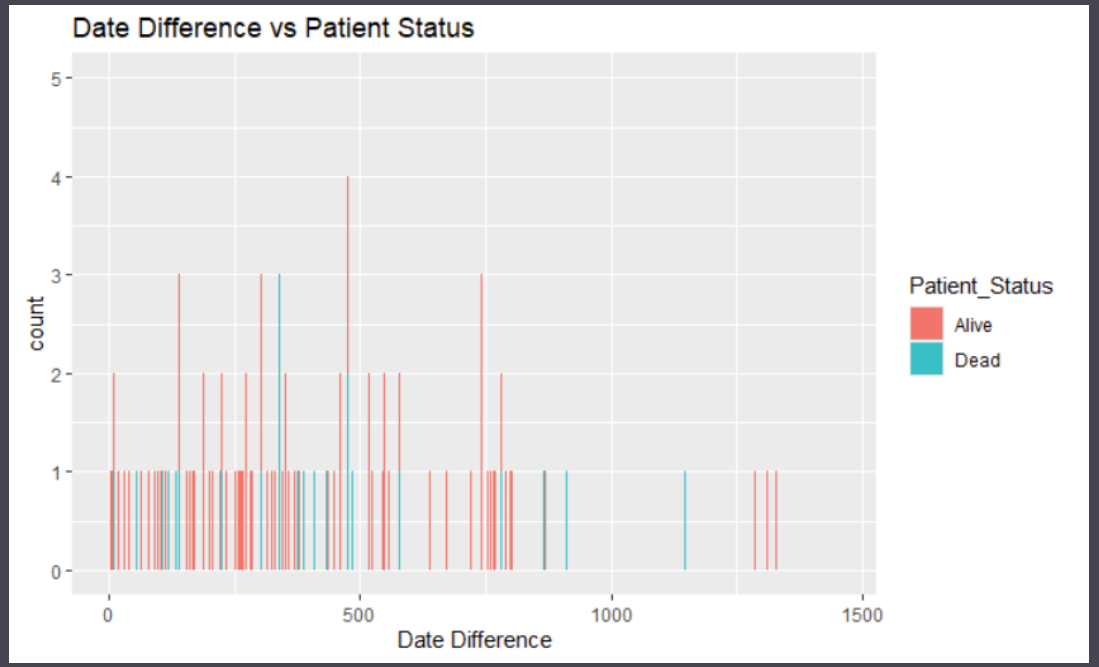
Covariance


Age vs. Gender
count
15
10
5
0
40
60
80
Age
Gender
FEMALE
MALE

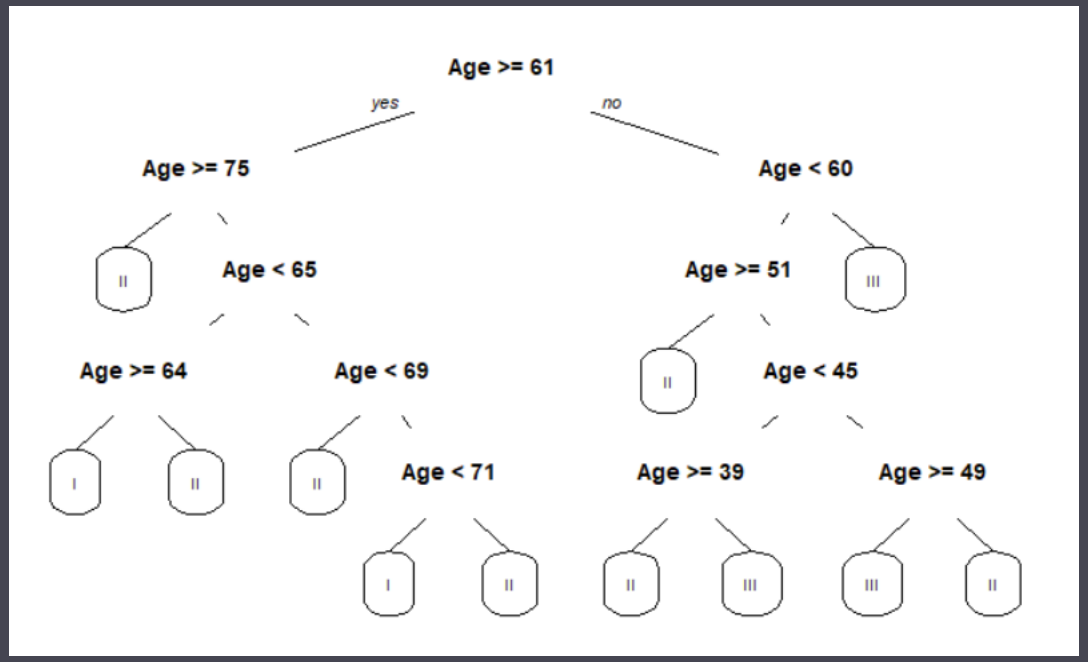

Age vs. Patient_Status and HER2
Alive
Dead
Count
10
5
0
40
60
80
Age
HER2
Negative
Positive

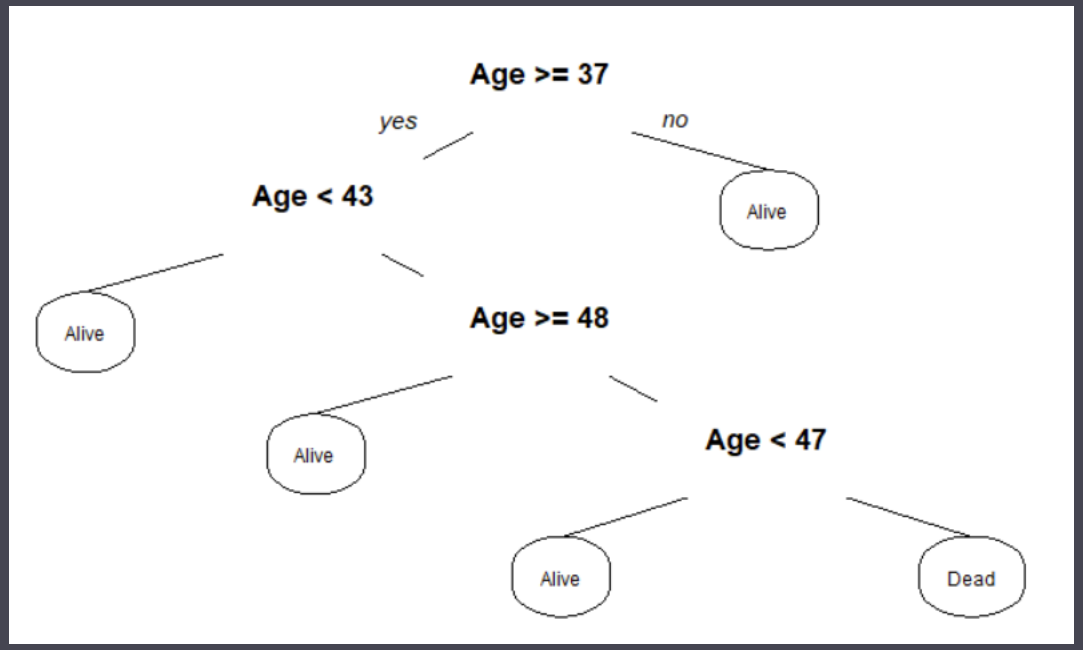

Tumor Stage vs. Histology
Infiltrating Ductal Carcinoma
Infiltrating Lobular Carcinoma
Mucinous Carcinoma
Count
125
100
75
50
25
0
I
II
III
Tumor Stage


Surgery type vs Patient Status
count
100
75
50
25
0
Lumpectomy
Modified Radical Mastectomy
Other
Simple Mastectomy
Surgery Type
Patient_Status
Alive
Dead


Date Difference vs Patient Status
count
5
4
3
2
1
0
0
500
1000
1500
Date Difference
Patient_Status
Alive
Dead

Decision Trees





Classification

Binarized dataset with Tumour_Stage as the class

Patient_ID	Gender	ER	PR	HER2	Patient_Status	Tumour_Stage
TCGA-D8-A1XD	1	0	0	1	1	III
TCGA-EW-A1OX	1	0	0	1	0	II
TCGA-A8-A079	1	0	0	1	1	III
TCGA-D8-A1XR	1	0	0	1	1	II
TCGA-BH-A0BF	1	0	0	1	0	II
...	...	...	...	...	...	...
TCGA-AN-A04A	1	0	0	0	0	III
TCGA-A8-A085	0	0	0	1	0	II
TCGA-A1-A0SG	1	0	0	1	0	II
TCGA-A2-A0EU	1	0	0	0	0	I
TCGA-B6-A40B	1	0	0	1	0	I

Tumour_Stage changed to early (stage 1 and 2) vs late (stage 3 and 4)

Patient_ID	Gender	ER	PR	HER2	Patient_Status	Tumour_Stage
TCGA-D8-A1XD	1	0	0	1	1	Late
TCGA-EW-A1OX	1	0	0	1	0	Early
TCGA-A8-A079	1	0	0	1	1	Late
TCGA-D8-A1XR	1	0	0	1	1	Early
TCGA-BH-A0BF	1	0	0	1	0	Early
...	...	...	...	...	...	...
TCGA-AN-A04A	1	0	0	0	0	Late
TCGA-A8-A085	0	0	0	1	0	Early
TCGA-A1-A0SG	1	0	0	1	0	Early
TCGA-A2-A0EU	1	0	0	0	0	Early
TCGA-B6-A40B	1	0	0	1	0	Early

Patient_Status	Tumour_Stage HER2	Early	Late
0	0	2	2
	1	46	16
1	0	14	11
	1	181	49

Patient_Status	Tumour_Stage Gender	Early	Late
0	0	1	0
	1	47	18
1	0	2	1
	1	193	59

HER2	Tumour_Stage Gender	Early	Late
0	1	16	13
1	0	3	1
	1	224	64

Regression

there is no clear, linear relationship between the variables you are trying to study

Cluster Analysis

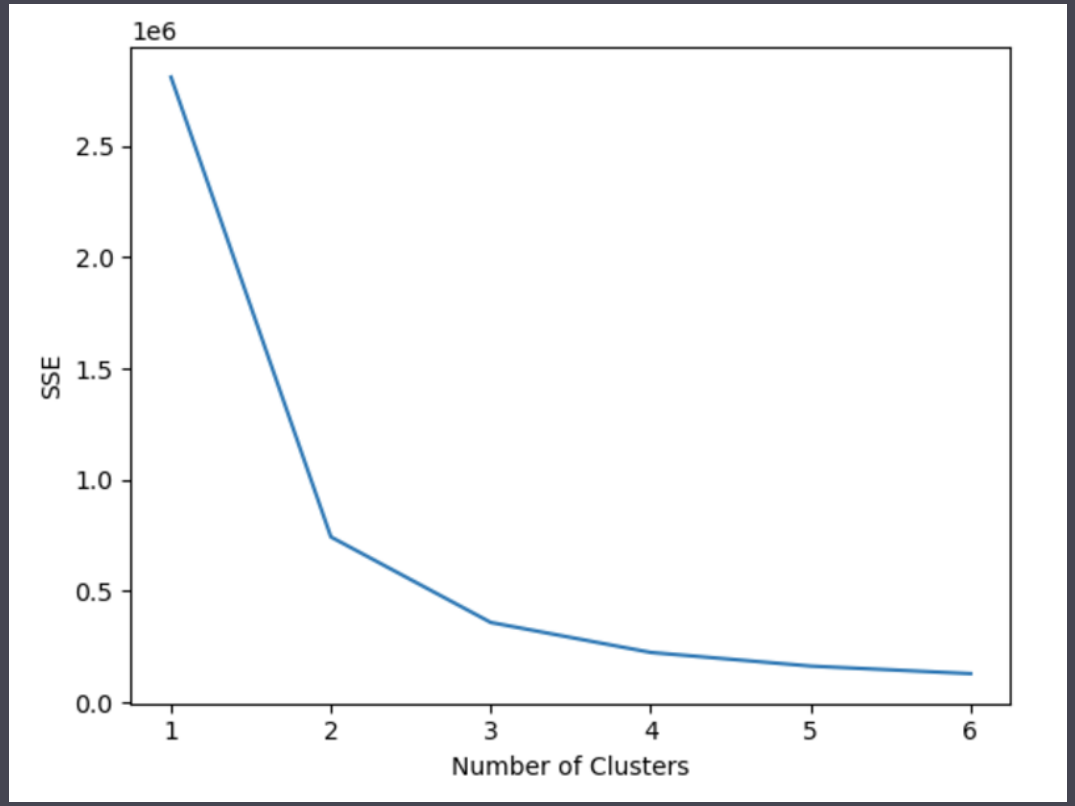
Patient_ID	Age	Protein1	Protein2	Protein3	Protein4
TCGA-D8-A1XD	36	0.080353	0.42638	0.54715	0.273680
TCGA-EW-A1OX	43	-0.420320	0.57807	0.61447	-0.031505
TCGA-A8-A079	69	0.213980	1.31140	-0.32747	-0.234260
TCGA-D8-A1XR	56	0.345090	-0.21147	-0.19304	0.124270
TCGA-BH-A0BF	56	0.221550	1.90680	0.52045	-0.311990
...	...	...	...	...	...
TCGA-AN-A04A	36	0.231800	0.61804	-0.55779	-0.517350
TCGA-A8-A085	44	0.732720	1.11170	-0.26952	-0.354920
TCGA-A1-A0SG	61	-0.719470	2.54850	-0.15024	0.339680
TCGA-A2-A0EU	79	0.479400	2.05590	-0.53136	-0.188480
TCGA-B6-A40B	76	-0.244270	0.92556	-0.41823	-0.067848

K-means Clustering

Patient_ID	Cluster ID
TCGA-D8-A1XD	1
TCGA-EW-A1OX	1
TCGA-A8-A079	1
TCGA-D8-A1XR	1
TCGA-BH-A0BF	1
...	...
TCGA-AN-A04A	0
TCGA-A8-A085	0
TCGA-A1-A0SG	0
TCGA-A2-A0EU	0
TCGA-B6-A40B	0

	Unnamed: 0	Age	Protein1	Protein2	Protein3	Protein4
0	241.0	58.627329	-0.001666	1.008145	-0.054543	-0.001991
1	80.5	59.125000	-0.051448	0.900647	-0.132257	0.019685

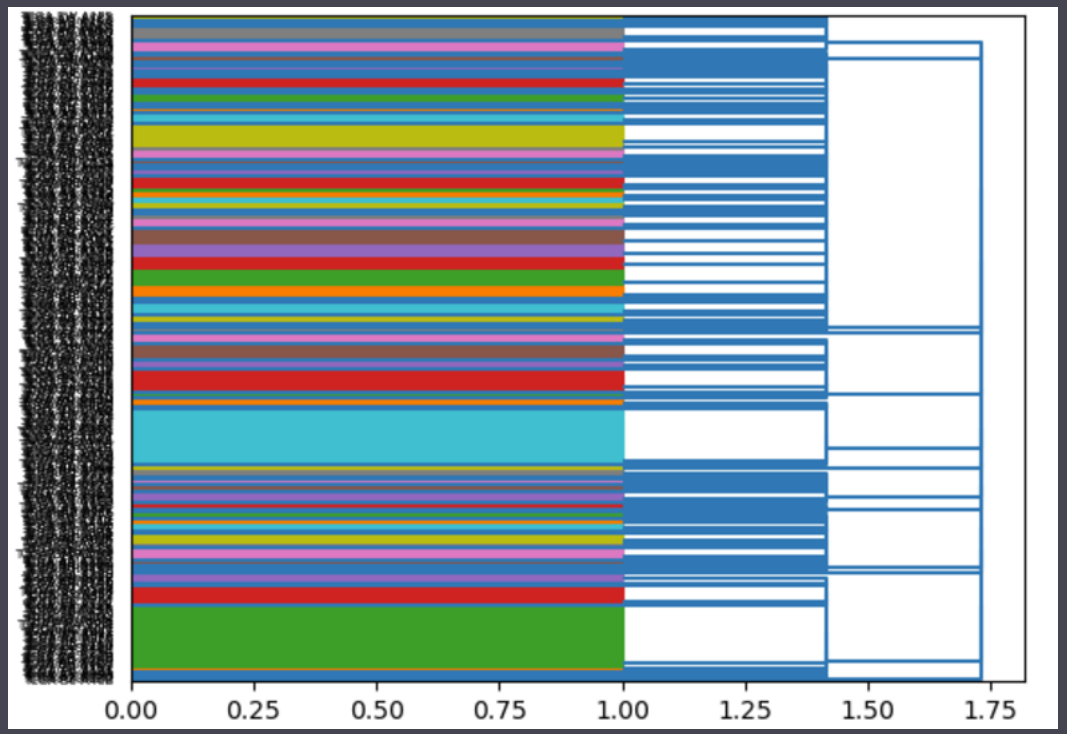
Estimate number of clusters

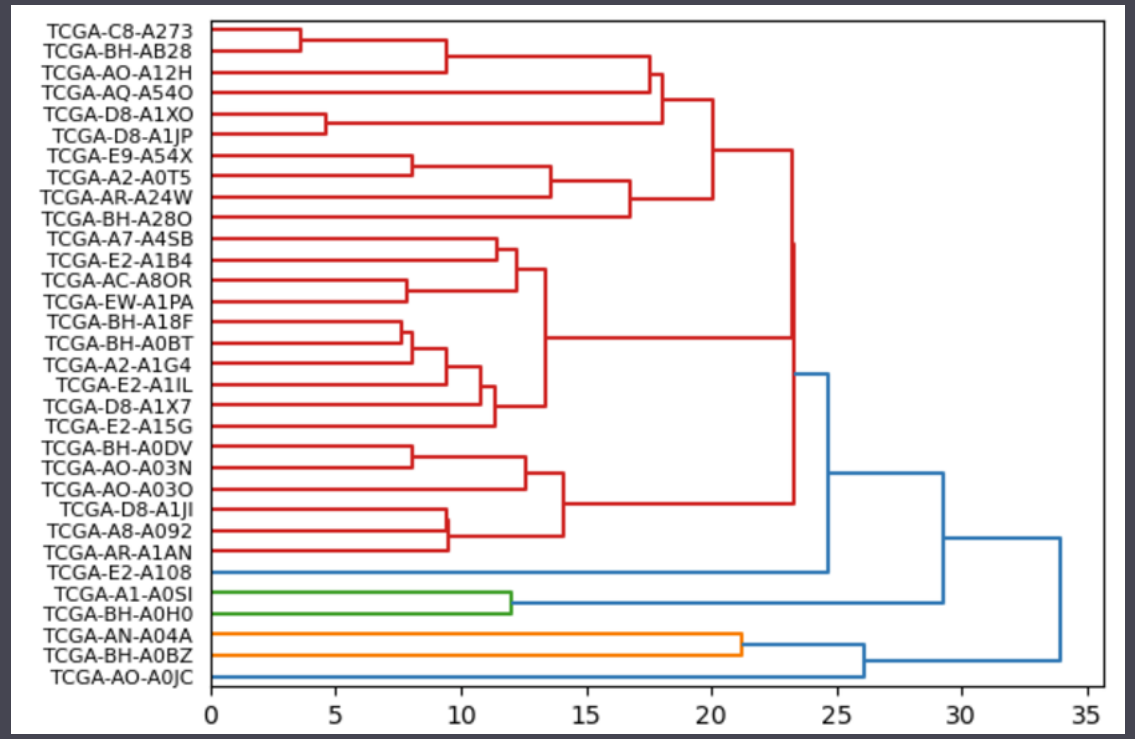


Hierarchical Clustering

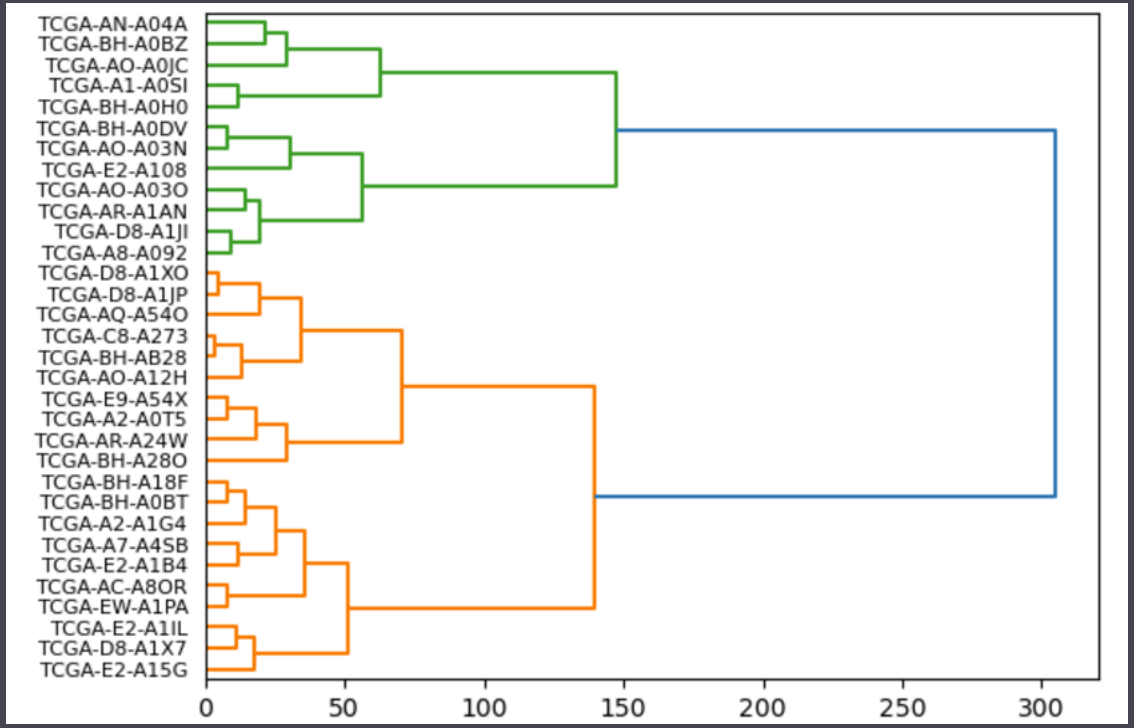
Patient_ID	Gender	ER	PR	HER2	Patient_Status	Tumour_Stage
TCGA-D8-A1XD	1	0	0	1	1	III
TCGA-EW-A1OX	1	0	0	1	0	II
TCGA-A8-A079	1	0	0	1	1	III
TCGA-D8-A1XR	1	0	0	1	1	II
TCGA-BH-A0BF	1	0	0	1	0	II
...	...	...	...	...	...	...
TCGA-AN-A04A	1	0	0	0	0	III
TCGA-A8-A085	0	0	0	1	0	II
TCGA-A1-A0SG	1	0	0	1	0	II
TCGA-A2-A0EU	1	0	0	0	0	I
TCGA-B6-A40B	1	0	0	1	0	I

Single Link (MIN)

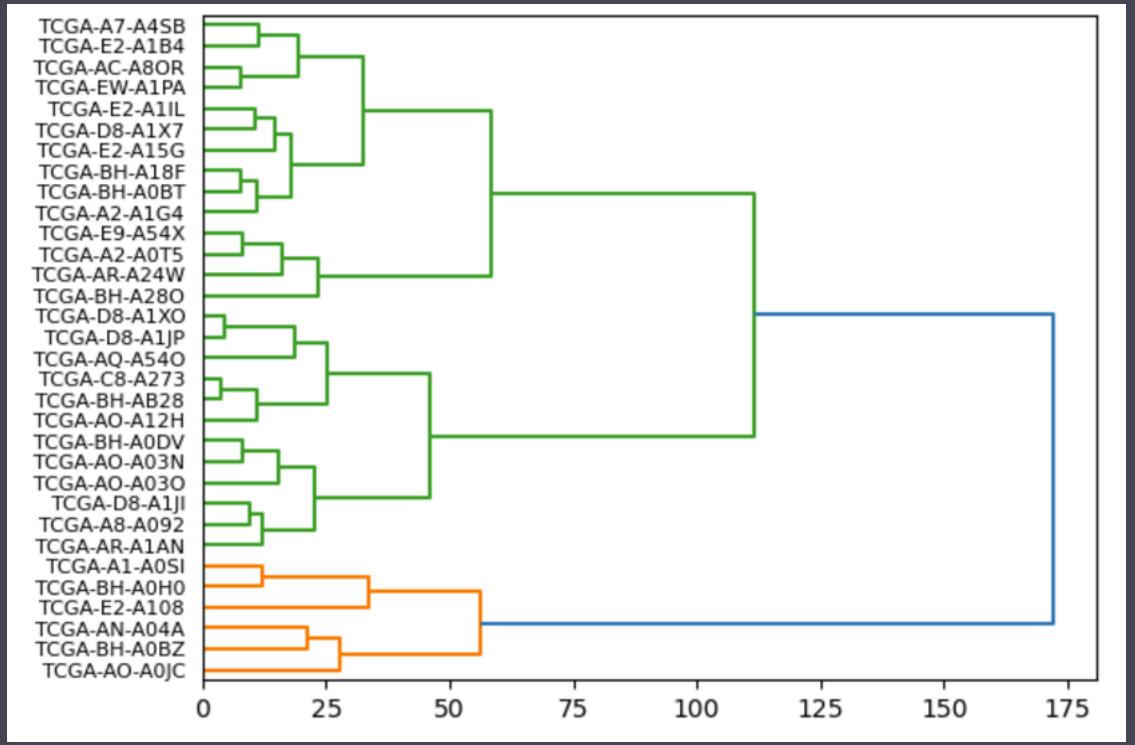




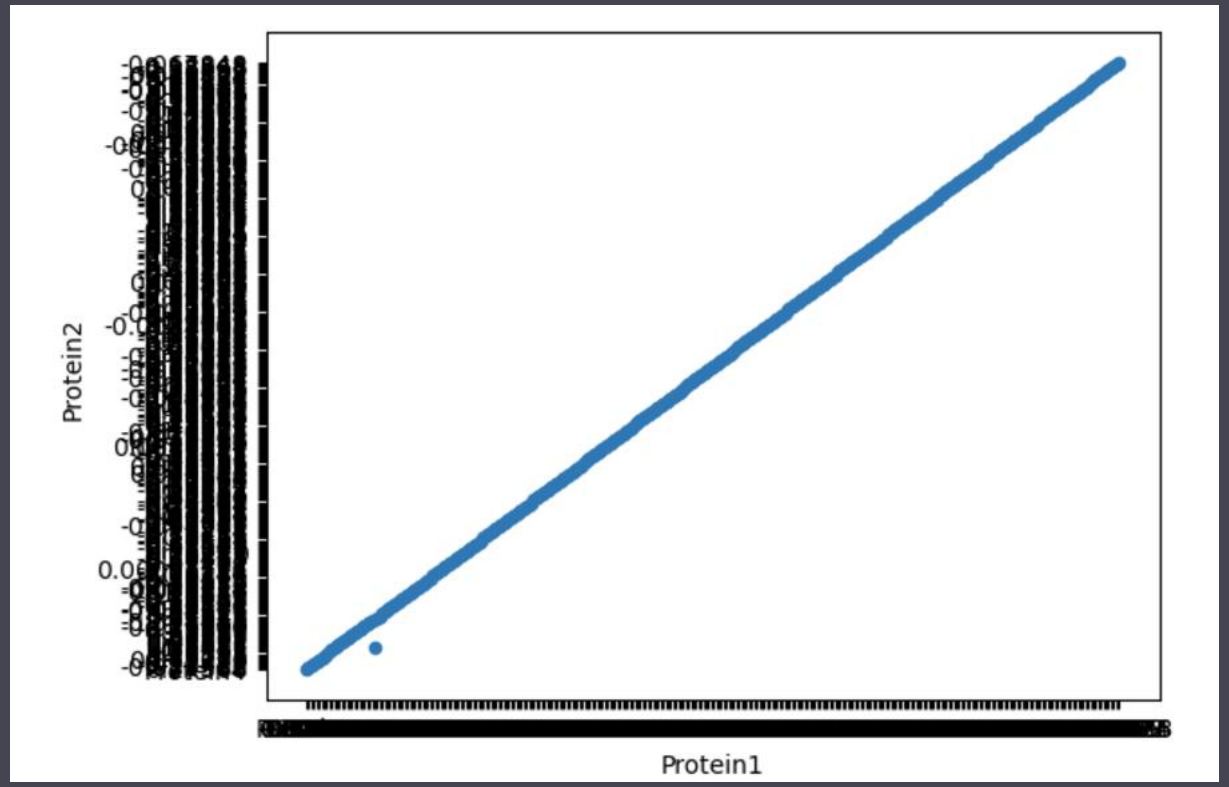
Complete Link (MAX)



Group Average



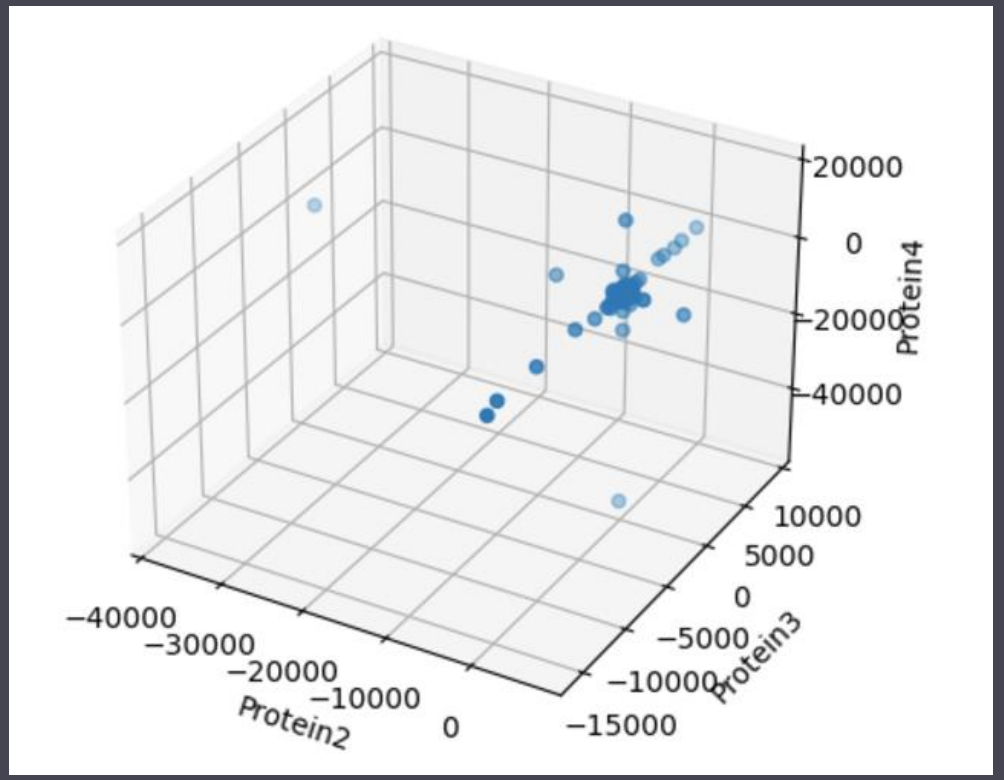
Density-Based Clustering



Anomaly Detection

Date_of_Surgery	Unnamed: 0	Protein2	Protein3	Protein4
17-01-15	1	0.42638	0.54715	0.273680
17-04-26	2	0.57807	0.61447	-0.031505
17-09-08	3	1.31140	-0.32747	-0.234260
17-01-25	4	-0.21147	-0.19304	0.124270
17-05-06	5	1.90680	0.52045	-0.311990

Plot distribution



Mean and Covariance Matrix

```
Unnamed: 0      1.983468
Protein2     -152.067681
Protein3     -157.005234
Protein4             inf
dtype: float64
            Unnamed: 0      Protein2      Protein3  Protein4
Unnamed: 0   47.521043  1.443255e+02 -2.312398e+01       NaN
Protein2    144.325485  4.929066e+06 -1.375878e+04       NaN
Protein3    -23.123983 -1.375878e+04  2.630912e+06       NaN
Protein4           NaN           NaN           NaN       NaN
```

Date_of_Surgery	Unnamed: 0	Protein2	Protein3	Protein4	Anomaly score
17-04-26	98.016532	187.643927	169.308990	-111.511619	NaN
17-09-08	48.016532	278.926019	3.712152	643.564514	NaN
17-01-25	31.349865	35.942166	115.954146	-153.047896	NaN
17-05-06	23.016532	-849.620502	-212.602101	-351.058180	NaN
17-09-18	18.016532	142.486183	45.988806	-113.790506	NaN

Association Rules

Only quantitative variables: Gender, Tumour_Stage, Histology, HER2, Surgery_Type, Patient_Status

1. Histology=Infiltrating Ductal Carcinoma 233 ==> Gender=FEMALE 231 <conf:(0.99)> lift:(1) lev:(0) [0] conv:(0.93)
2. Histology=Infiltrating Ductal Carcinoma HER2 status=Negative 212 ==> Gender=FEMALE 210 <conf:(0.99)> lift:(1) lev:(0) [0] conv:(0.85)
3. Patient_Status=Alive 255 ==> Gender=FEMALE 252 <conf:(0.99)> lift:(1) lev:(0) [0] conv:(0.76)
4. HER2 status=Negative Patient_Status=Alive 230 ==> Gender=FEMALE 227 <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.69)
5. HER2 status=Negative 305 ==> Gender=FEMALE 301 <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.73)
6. Gender=FEMALE 330 ==> HER2 status=Negative 301 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.96)
7. Histology=Infiltrating Ductal Carcinoma 233 ==> HER2 status=Negative 212 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.92)
8. Gender=FEMALE Histology=Infiltrating Ductal Carcinoma 231 ==> HER2 status=Negative 210 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.91)
9. Patient_Status=Alive 255 ==> HER2 status=Negative 230 <conf:(0.9)> lift:(0.99) lev:(-0.01) [-2] conv:(0.85)
10. Histology=Infiltrating Ductal Carcinoma 233 ==> Gender=FEMALE HER2 status=Negative 210 <conf:(0.9)> lift:(1) lev:(0) [0] conv:(0.96)

Neural Network

Tumour_Stage, Histology, Surgery_type, Patient_Status

Surgery Type

```
=== Summary ===

Correctly Classified Instances         251               78.1931 %
Incorrectly Classified Instances        70               21.8069 %
Kappa statistic                          0.0263
Mean absolute error                      0.3174
Root mean squared error                  0.4219
Relative absolute error                 96.5449 %
Root relative squared error            104.3914 %
Total Number of Instances              321
Ignored Class Unknown Instances                13

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973     0.955     0.797      0.973     0.876      0.042    0.526     0.764     Alive
                 0.045     0.027     0.300      0.045     0.079      0.042    0.528     0.231     Dead
Weighted Avg.    0.782     0.764     0.695      0.782     0.712      0.042    0.527     0.654

=== Confusion Matrix ===

   a   b   <-- classified as
 248   7 |   a = Alive
  63   3 |   b = Dead
```

Histology

```
=== Summary ===

Correctly Classified Instances         206               61.6766 %
Incorrectly Classified Instances       128               38.3234 %
Kappa statistic                          0.04
Mean absolute error                      0.2957
Root mean squared error                  0.4083
Relative absolute error                 99.6719 %
Root relative squared error            106.4824 %
Total Number of Instances              334

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.785     0.752     0.707       0.785   0.744      0.036    0.507     0.705     Infiltrating Ductal Carcinoma
                 0.000     0.000     ?           0.000   ?          ?        0.619     0.100     Mucinous Carcinoma
                 0.258     0.212     0.307       0.258   0.280      0.049    0.510     0.276     Infiltrating Lobular Carcinoma
Weighted Avg.    0.617     0.581     ?           0.617   ?          ?        0.512     0.569

=== Confusion Matrix ===

   a   b   c   <-- classified as
 183   0  50 |   a = Infiltrating Ductal Carcinoma
  10   0   2 |   b = Mucinous Carcinoma
  66   0  23 |   c = Infiltrating Lobular Carcinoma
```

Tumor Stage

```
=== Summary ===

Correctly Classified Instances         188               56.2874 %
Incorrectly Classified Instances       146               43.7126 %
Kappa statistic                          0.0852
Mean absolute error                      0.3652
Root mean squared error                  0.4422
Relative absolute error                 93.5143 %
Root relative squared error            100.205  %
Total Number of Instances              334

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.247     0.095     0.455       0.247   0.320      0.193    0.662     0.381     III
               0.889     0.834     0.581       0.889   0.703      0.079    0.580     0.634     II
               0.000     0.004     0.000       0.000   0.000      -0.027   0.569     0.230     I
Weighted Avg.  0.563     0.496     0.439       0.563   0.475      0.086    0.598     0.495

=== Confusion Matrix ===

   a   b   c   <-- classified as
  20  61   0 |   a = III
  20 168   1 |   b = II
   4  60   0 |   c = I
```

Results/Suggestions

- Age decision tree
- Cluster analysis, Regression, and Neural Network
- Suggestions: Look closer at HER2 protein in regards to Tumour Stage and Patient Status
- Plans: cluster analysis and anomalies