Breast Cancer Data Review

Alyssa Wheeler

Introduction:

This dataset provides the data from a group of breast cancer patients that had surgery to remove their tumors. The goal is to find any connections between variables such as age and tumour stage that could help in patient outcome or risk. This project is aimed at providing helpful information to women or others who have breast cancer related concerns. The source of the dataset can be found at https://www.kaggle.com/datasets/amandam1/breastcancerdataset.
The CSV is called "BRCA" and is located on Kaggle. It can be downloaded at the tab on the top right.

Data Dictionary Table:

Attribute	Type	Description
Patient_ID	Arbitrary	Unique identifier id of a patient
Age (29-90)	Continuous	Age at diagnosis (Years)
Gender	Categorical	Male/Female
Protein1, Protein2, Protein3, Protein4	Continuous	Expression levels (undefined units)
Tumour_Stage	Ranked	I, II, III
Histology	Categorical	Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, Mucinous Carcinoma
ER status	Boolean	Estrogen receptor hormone to test if breast cancer cells are Positive/Negative
PR status	Boolean	Progesterone receptor hormone to test if breast cancer cells are Positive/Negative
HER2 status	Boolean	Human epidermal growth factor receptor 2 protein to test if breast cancer cells are Positive/Negative
Surgery_type	Categorical	Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other
Date_of_Surgery	Chronological	Date on which surgery was performed (in YY-MM-DD)
Date_of_Last_Visit	Chronological	Date of last visit (in YY-MM-DD) [null, in case the patient didn't visited again after the surgery]
Patient_Status	Categorical	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]
DateDiff	Continuous	Difference in days between the date of surgery and the last visit

Example of 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

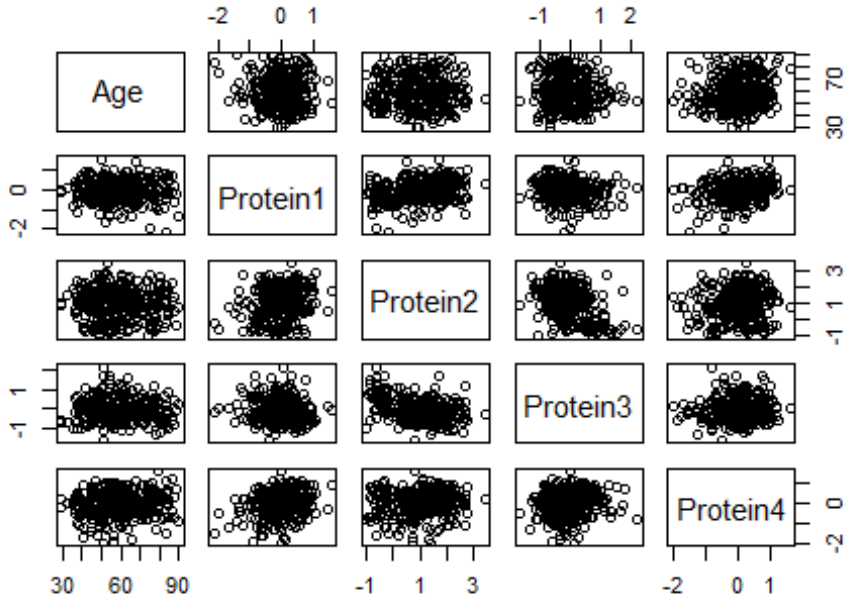
ER <chr>	PR <chr>	HER2 <chr>	Surgery_type <chr>	Date_of_Surgery <chr>	Date_of_Last_Visit <chr>	Patient_Status <chr>	DateDiff <dbl>
Positive	Positive	Negative	Modified Radical Mastectomy	17-01-15	17-06-19	Alive	155
Positive	Positive	Negative	Lumpectomy	17-04-26	18-11-09	Dead	562
Positive	Positive	Negative	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

Data Cleaning and Manipulation:

Preprocessing steps summarized (RStudio): There was no merging needed. There was a variable created called DateDiff that is the difference in days between the last visit and date of surgery. Some variable names needed to be changed to remove spaces. Patient_Status needed to have the NA's removed. ER and PR hormones are all positive, so they may not be looked at because they don't convey helpful information that distinguishes one patient from another.

Data Exploration and Visualization:

Scatter Plot Matrix for Numeric Attributes in Dataset



The scatterplot matrix showed that there were a few attributes that could be looked closer at for a relationship with each other. This is indicated by a more linear pattern between continuous attributes. These ones include: Age and Protein1, Age and Protein3, Age and Protein4, Protein1 and Protein2, and Protein2 and Protein3.

Figure 1: Age vs Gender

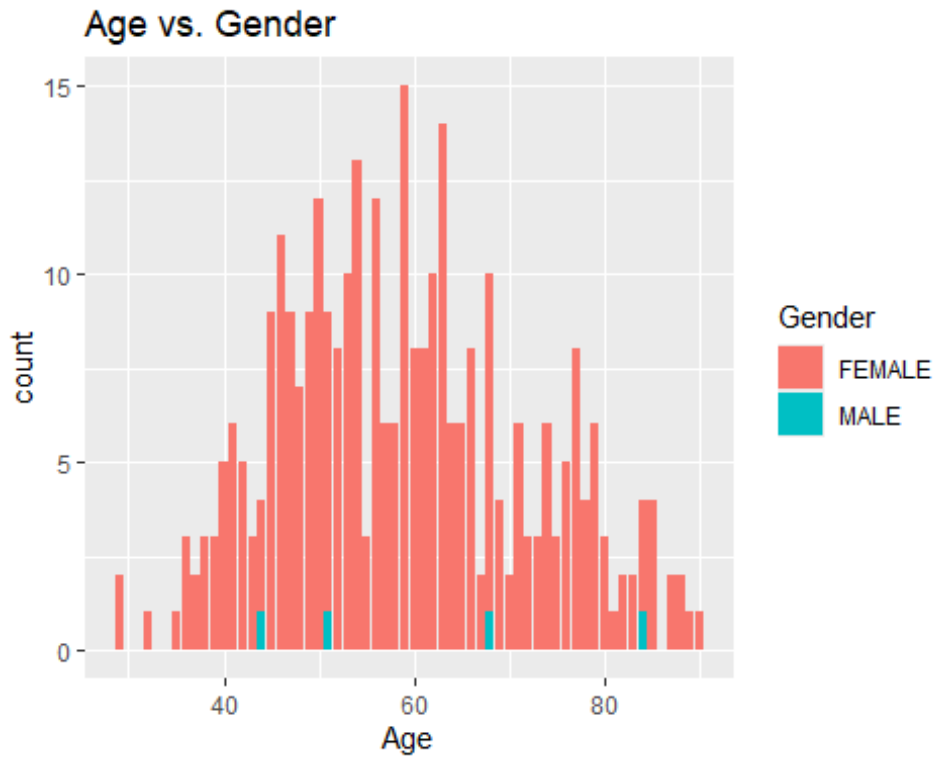


Figure 2: Age vs Patient_Status and HER2

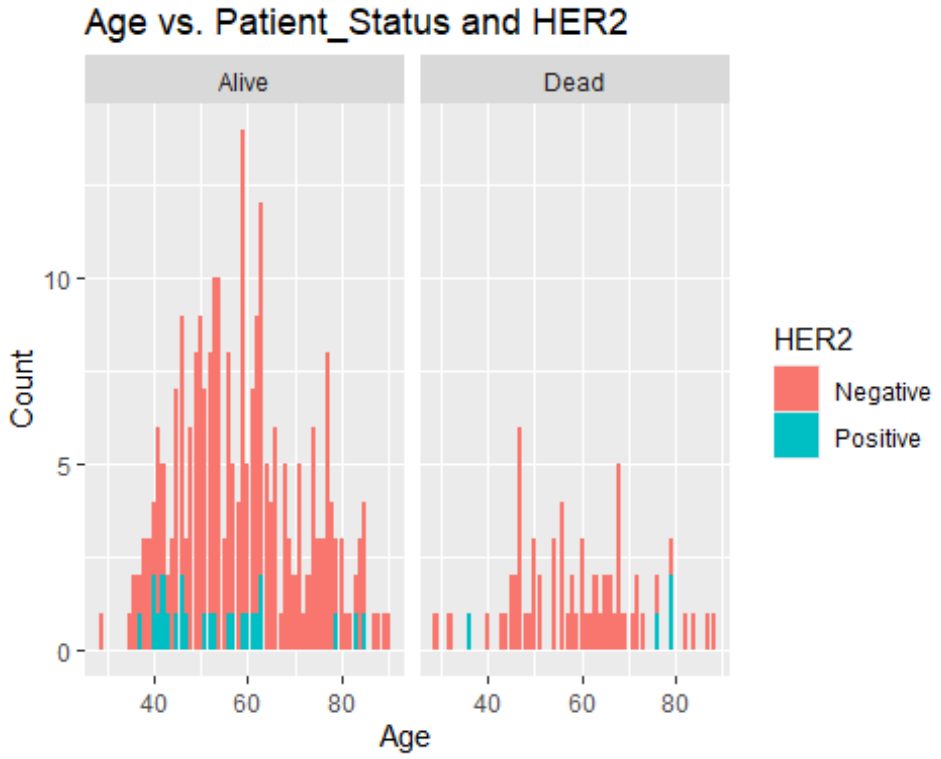


Figure 3: Tumour_Stage vs Histology

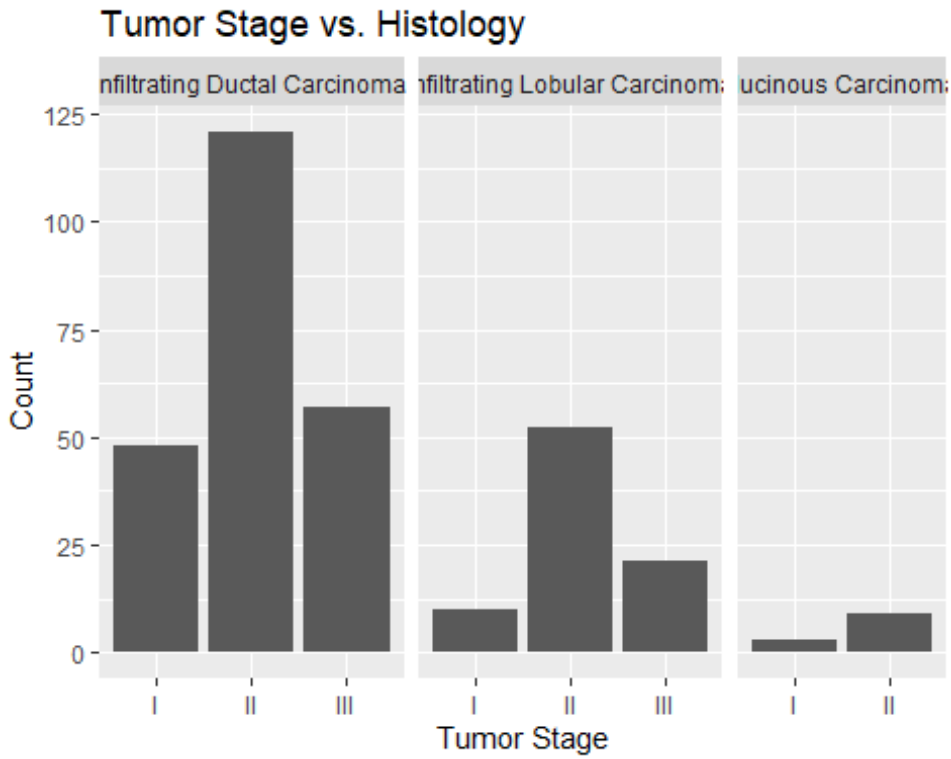


Figure 4: Surgery_Type vs Patient_Status

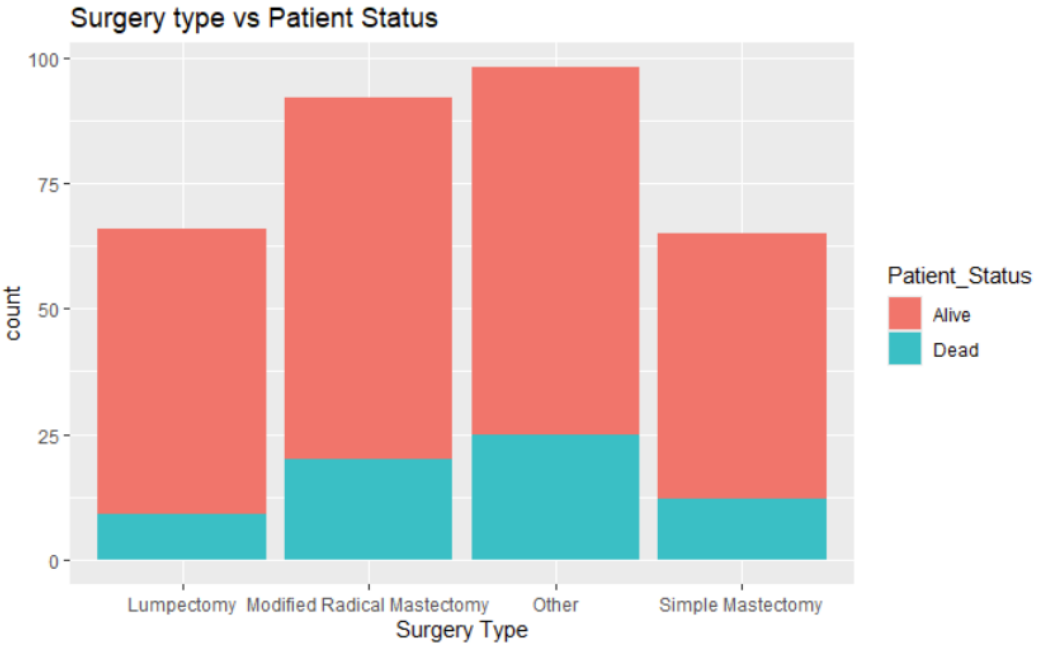
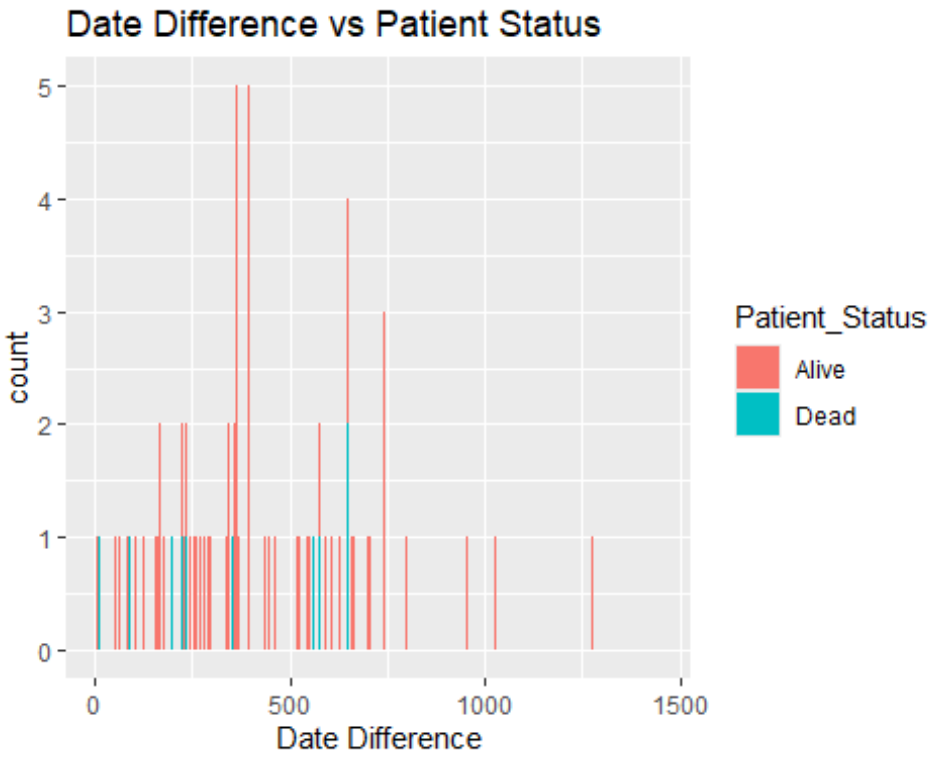


Figure 5: DateDiff vs Patient_Status



Preliminary Observations:
Both hormones ER and PR were positive for all patients. There are few Males compared to Females in the dataset. There are a wide range of ages. Protein1-4 are not named which is unhelpful but still useful for some comparisons. There are some NA's in the Date_of_Last_Visit variable, but this means the person didn't come back after surgery. These were left in the dataset. There doesn't seem to be any usefulness between the DateDiff and Patient_Status, but there may be other comparisons with this that could be helpful.

Data Mining:

Decision Tree (RStudio)

Figure 6: Tumour Stage Decision Tree

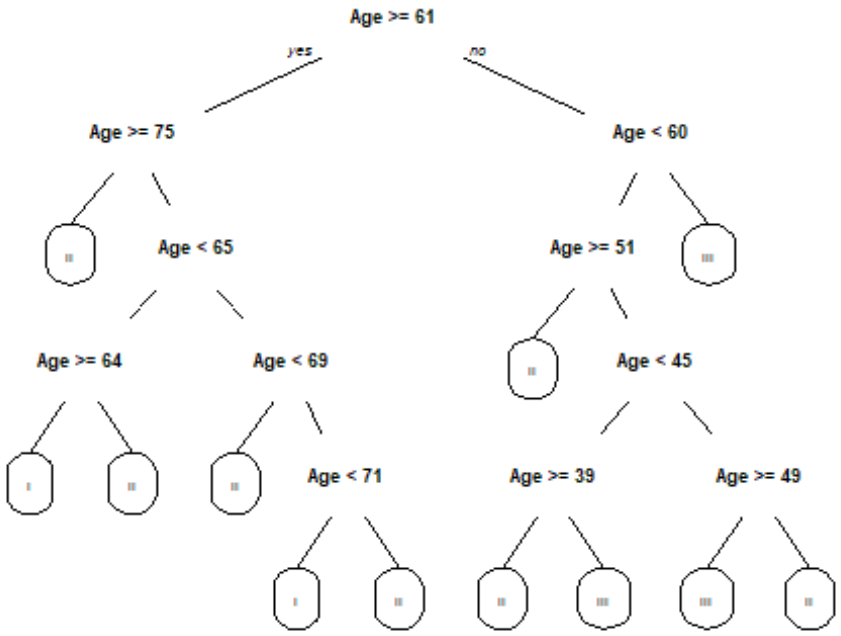
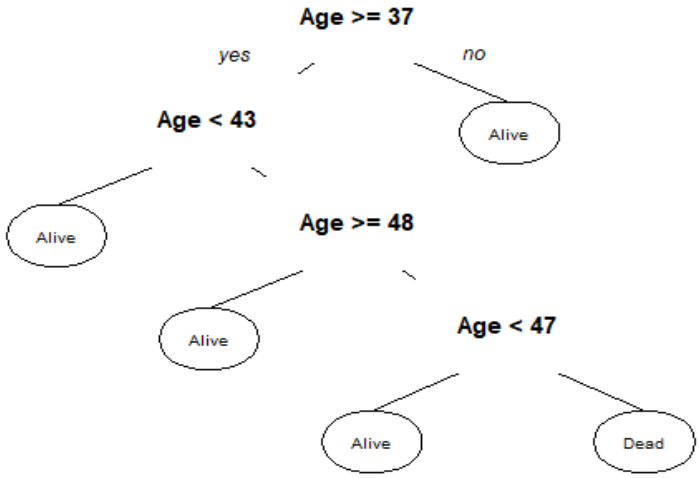


Figure 7: Patient Status Decision Tree



The tumour stage decision tree shows pretty mixed results with which age determines what stage tumor. This makes some sense since it may be more of a factor of when a person first realizes and catches the cancer. The second decision tree shows a better patient status outcome if the patient is less than 37 years old, and a worse outcome if the patient is greater than 47 years old. It would be interesting to look more into the different factors that affect this age difference.

Classification (Python)

Table 1: Binarized Dataset

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

Classification is the task of predicting a nominal-valued attribute (class label) based on the values of other attributes (predictor variables). To do this, I looked at only the attributes that could be binarized. Tumour_Stage needed to have only two options. The stage 1 and 2 were changed to “Early” and the stage 3 and 4 were changed to “Late.” The index is the Patient_ID, and the class that we are trying to predict is the Tumour_Stage. I applied Pandas cross-tabulation to examine the relationship between the Patient_Status and HER2 attributes, Patient_Status and Gender, and HER2 and Gender with respect to the Tumour_Stage.

Table 2: Patient_Status and HER2 with respect to Tumour_Stage

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

Table 3: Patient_Status and Gender with respect to Tumour_Stage

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

Table 4: HER2 and Gender with respect to Tumour_Stage

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

Looking at Table 2, this table compares Patient_Status to HER2, Patient status of 0 is dead while negative HER2 is 1. For both Patient_Status, the HER2 when negative was larger. For Table 3, the Gender being female was larger for both Patient_Status. In Table 4, Gender being female was also larger for both negative and positive HER2 with Negative HER2 being largest. There isn't much difference in Gender (very few males) in the dataset, so it isn't as helpful to look at those and makes sense that in the two tables using Gender, the 0 which represents male has low values.

Regression

There was not enough of a linear relationship between the variables for regression to be helpful with this dataset.

Cluster Analysis (Python)

Cluster analysis seeks to partition the input data into groups of closely related instances so that instances that belong to the same cluster are more similar to each other than to instances that belong to other clusters. The k-means clustering algorithm represents each cluster by its corresponding cluster centroid.

Table 5: Clustering Data

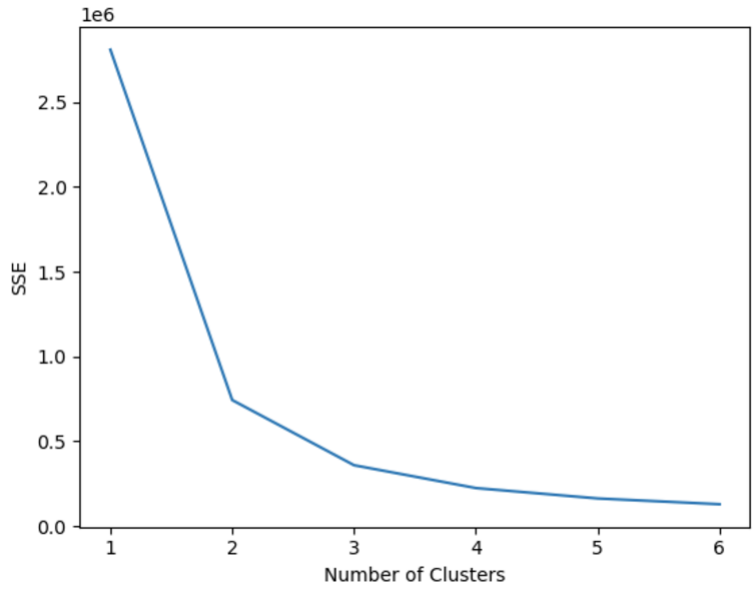
Patient_ID	Age	Protein1	Protein2	Protein3	Protein4
TCGA-D8-A1XD	36	0.080353	0.42638	0.54715	0.273680
TCGA-EW-A1OX	43	-0.420320	0.57807	0.61447	-0.031505
TCGA-A8-A079	69	0.213980	1.31140	-0.32747	-0.234260
TCGA-D8-A1XR	56	0.345090	-0.21147	-0.19304	0.124270
TCGA-BH-A0BF	56	0.221550	1.90680	0.52045	-0.311990
...	...	...	...	...	...
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

Table 6: K-Means Clustering

	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

Table 5 shows the data that was used for clustering, and Table 6 shows the results of clustering. As can be seen in Table 6, the dataset did not seem to cluster very well. There was not too much of a difference between the two clustering groups when looking at each attribute.

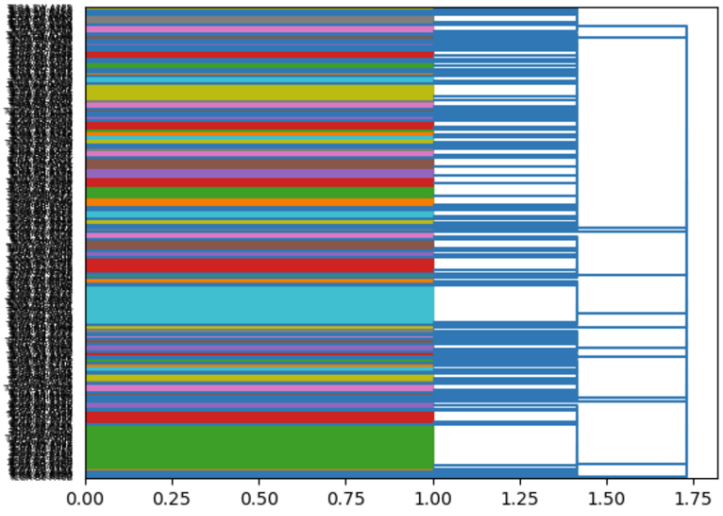
Figure 8: sum-of-squared errors (SSE)



To determine the number of clusters in the data, we can apply k-means with varying number of clusters from 1 to 6 and compute their corresponding sum-of-squared errors (SSE) as shown in Figure 8. The "elbow" in the plot of SSE versus number of clusters can be used to estimate the number of clusters and shows that for this dataset the number of clusters should be 2.

Hierarchical Clustering

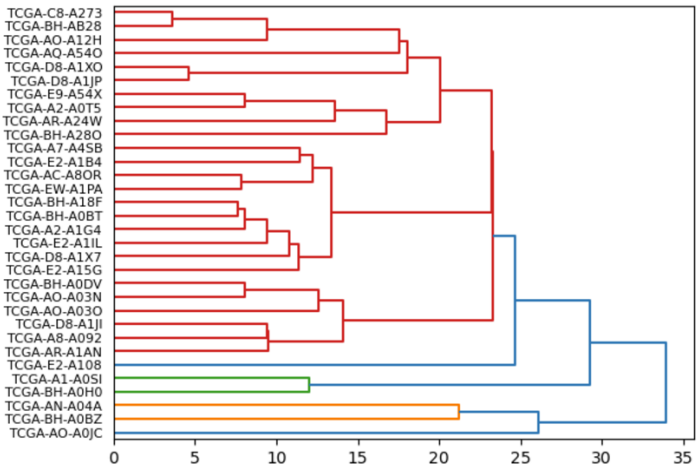
Figure 9: Single Link (MIN) without Sampling



For this clustering, I used the binarized part of the dataset and clustered based on Tumour_Stage. Single Link clustering uses the distance between two clusters as defined as the minimum distance between any pair of points from each cluster. Figure 9 shows single link clustering. Because there were so many patients in the study, as you can see from how

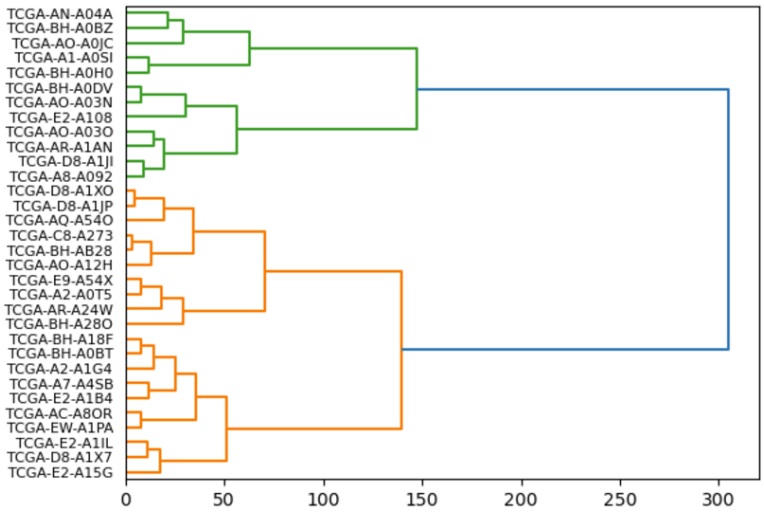
busy Figure 9 is, I did a random sampling of 10% of the dataset for the rest of the clustering. The sampled dataset single link clustering can be seen in Figure 10 below.

Figure 10: Single Link (MIN) with Sampling



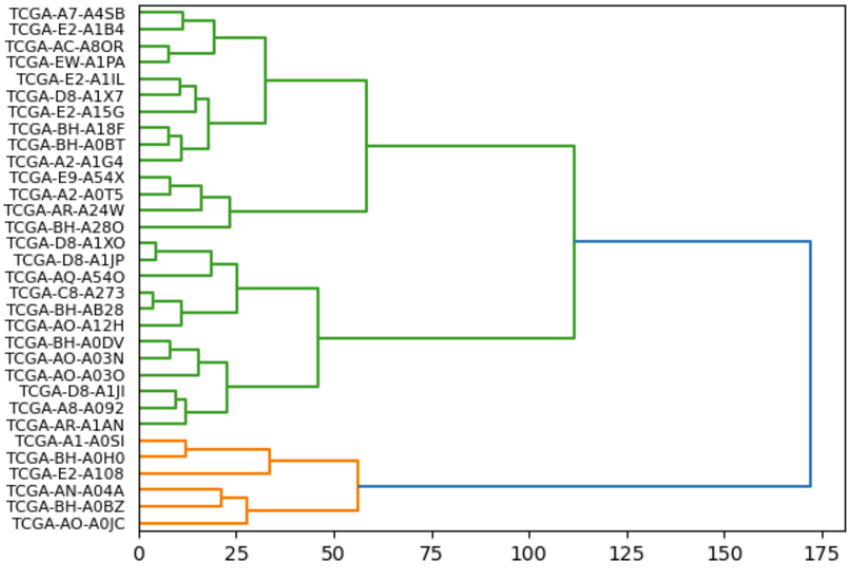
Although Single Link clustering seemed to work with the sampled dataset, it wasn't as helpful because I would need to go through each Patient_ID to really understand the clustering.

Figure 11: Complete Link (MAX)



Complete Link is where the distance between two clusters is defined as the maximum distance between any pair of data points, one from each cluster. Figure 11 shows this clustering.

Figure 12: Group Average



Group Average clustering is found from the average value of a specific variable within a group of data points that share similar characteristics. This can be seen in Figure 12. Both Figure 11 and 12 have the same problem as Figure 10 where there is clustering with the sampled dataset, but because the patient ID is used there is nothing to show why those Patient_IDs are clustering more than others without having to go through each Patient_ID.

Density-Based Clustering (Python)

Figure 13: Density-Based Clustering

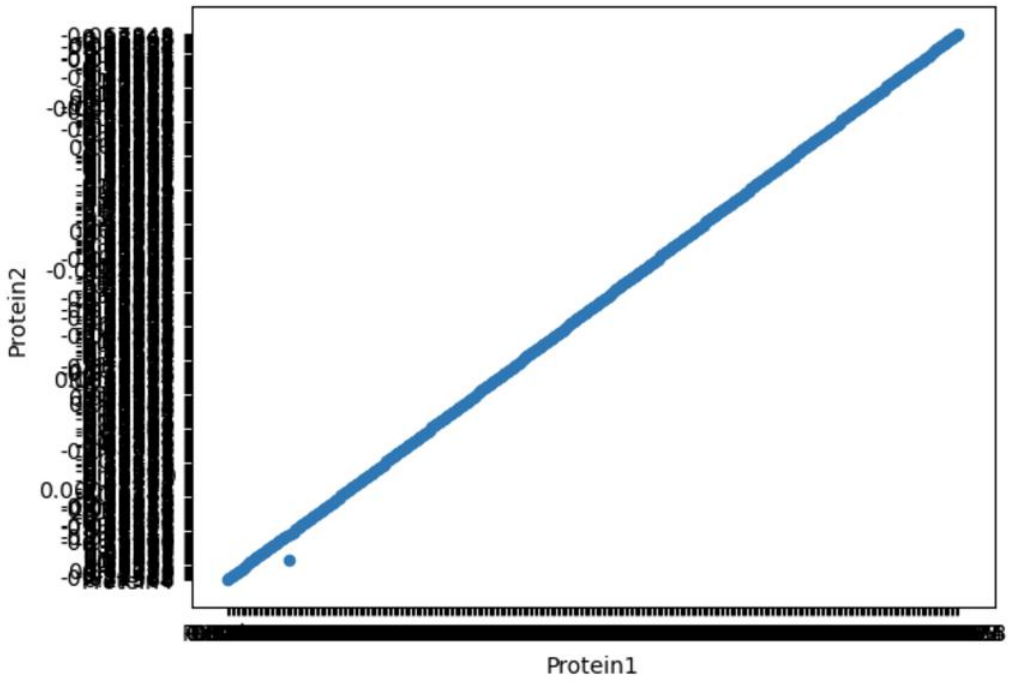


Figure 13 shows Density-Based clustering. This identifies clusters by grouping together points that are located in areas of high data density. This dataset did not lend itself to this type of clustering. All of the graphs when comparing any attributes looked similar to the one shown in Figure 13.

Anomaly Detection

Anomaly detection is the task of identifying instances whose characteristics differ significantly from the rest of the data. For this dataset, I looked at the date of surgery and Proteins 2,3, and 4.

Table 6: Anomaly Detection Dataset

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

Figure 14: Plot Distribution

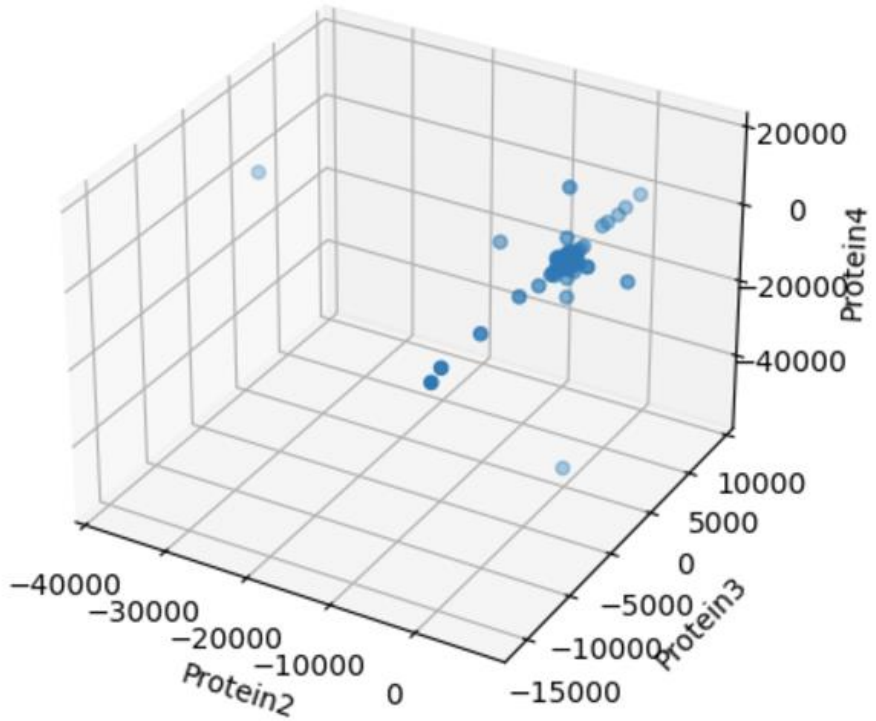


Figure 14 shows the plot distribution which shows a few points plotted far away from the others that could be looked at more closely.

Mean and Covariance Matrix (Python)

Table 7: Mean and Covariance

```
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
```

Table 8: Anomalies

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

The mean and covariance matrix assumes the data follows a multivariate Gaussian distribution. This data did not have that distribution, so this matrix could not be done as seen by the many "NaN" in Table 7 and Table 8.

Association Rules (Weka)

Only quantitative variables were used in the association rules: Gender, Tumour_Stage, Histology, HER2, Surgery_Type, Patient_Status.

Table 9: Association Rules

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)

Many of the rules, as seen in Table 9, made sense, but few were helpful. Many show that there is a relationship between Histology=Infiltrating Ductal Carcinoma, Gender=Female, and HER2 status=Negative. These three seem to have a strong association with each other.

Neural Network (Weka)

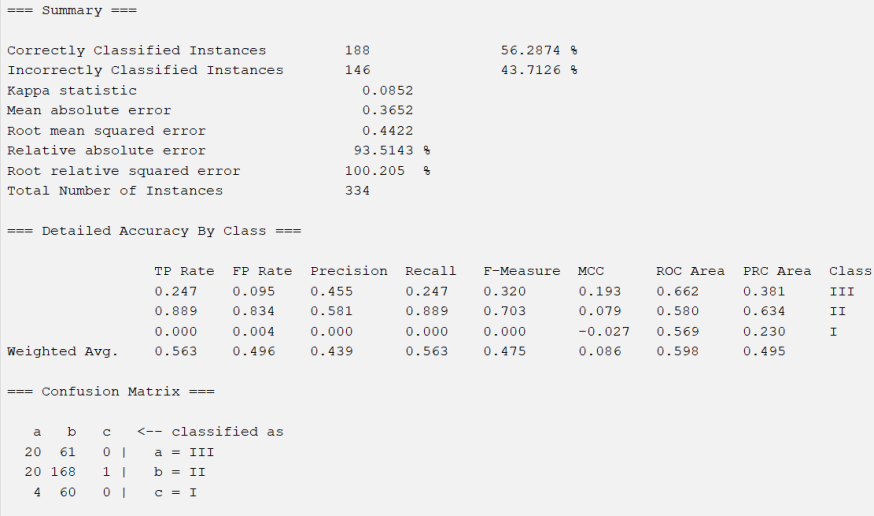
Figure 15: Surgery Type Neural Network

```
=== 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
```

Figure 16: Histology Neural Network

```
=== 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
```

Figure 17: Tumour Stage Neural Network



```
=== 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
```

A neural network is simulated with a perceptron or node that essentially takes the inputs and calculates an output based on weights of the inputs. Figures 15, 16, and 17 show different neural networks. Figure 15 was the most successful one, and it was based on the Surgery_Type. In this figure, the correctly classified instances were 78%. In Figures 16 and 17, the correctly classified instances were not very high meaning the neural network did not work very well.

Results

There were a few data mining techniques that proved to be helpful for this dataset. These were decision trees, classification, and association rules. The Age decision tree shows a better patient status outcome if the patient is less than 37 years old, and a worse outcome if the patient is greater than 47 years old. Classification showed HER2 negative for most patients regardless of Patient_Status. The association rules showed a relationship between Histology=Infiltrating Ductal Carcinoma, Gender=Female, and HER2 status=Negative. Each of these results could be indicators for potential risk for breast cancer or for signs that a patient may have a better or worse outcome with breast cancer. This dataset proved to have much bias in terms of type of data collected, and to make more conclusive results, more data would need to be collected that includes possibly other factors that differentiate patients better.