Breast Cancer Clinical

And Gene Expression Data Analysis

Cody Le

Varsha Sajja

Aaron Gregory

Evan Morton

DSC 424

Spring 2021

DePaul University

June 9, 2021

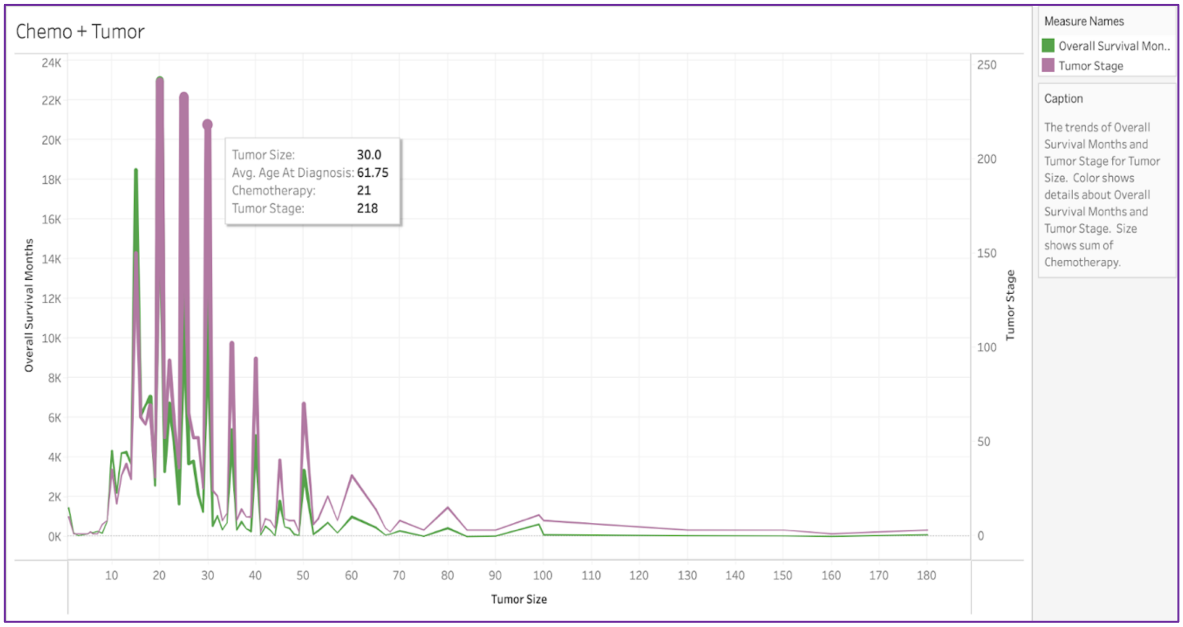
Table of Contents

Overview (Non-Technical Summary)

In 2020, breast cancer became the most common cancer worldwide surpassing lung cancer with more than 2.3 million women diagnosed according to the Breast Cancer Research Foundation (BCRF). Breast cancer is the most frequently diagnosed cancer in women and now represents one in four of all cancers in women (BCRF). Globally, in the last decade, diagnosis of breast cancer has increased by more than 20 percent, mortality has also increased by 14 percent, with 685,000 deaths related to breast cancer in 2020 (BCRF). Breast cancer patients with the same stage of the disease and the same clinical characteristics can have different treatment responses and overall survival. In addition, cancers are associated with genetic mutations that may affect the outcome of survival. An analysis of breast cancer patients' clinical and gene expression attributes as it relates to their survival time may bring better insights into the cancer prognosis and outcomes. Lastly, is the cause of death in patients solely from cancer? Analyzing the cause of death and probability of death from cancer may bring further insights to how patients are dying and what contributes to their death.

In this analysis, the breast cancer gene expression dataset will be explored which was acquired through the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, a project between Canada and the United Kingdom, and collected by Professor Carlos Caldas from the Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre, published in Nature Communications (Pereira et al., 2016). The dataset contains clinical variables and gene expression variables which after performing a dimensionality reduction through principal component analysis, the genomic variables were reduced significantly to capture only the most significant gene expressions for the analysis. The selected clinical variables contained numeric, binary, categorical, and ordinal variables and various techniques were performed on both clinical and genomic variables to explore their relationships including factor analysis, cluster analysis, correspondence analysis, and linear discriminant analysis.

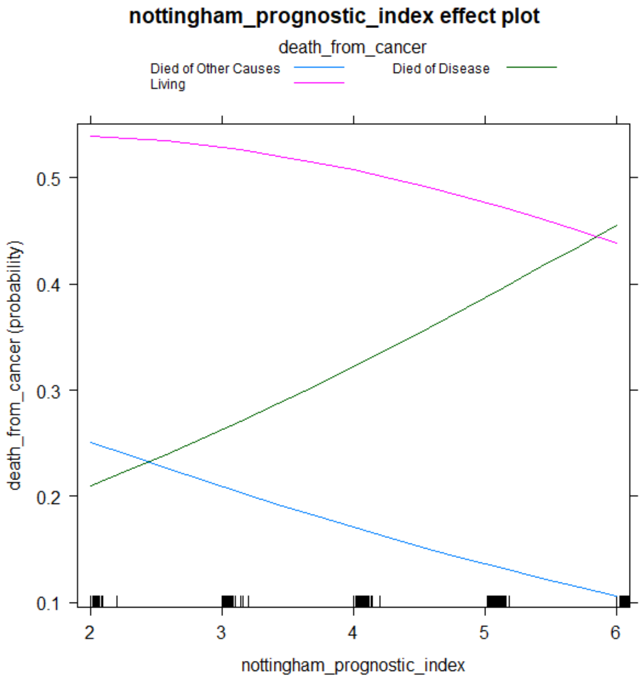
During the exploration of the clinical variables, overall survival years were negatively correlated with all other variables. Once cancer was detected, tumor size and tumor stage played a key role in the suggestion of therapies including chemotherapy and hormone therapy which can regulate cancer cells. The visualization of tumor size to overall survival years below shows that tumor stage is a significant variable. Ultimately, the stage of the tumor allowed for a better understanding of all other clinical variables as the stage of the cancer directly dictated the results of all other clinical variables.

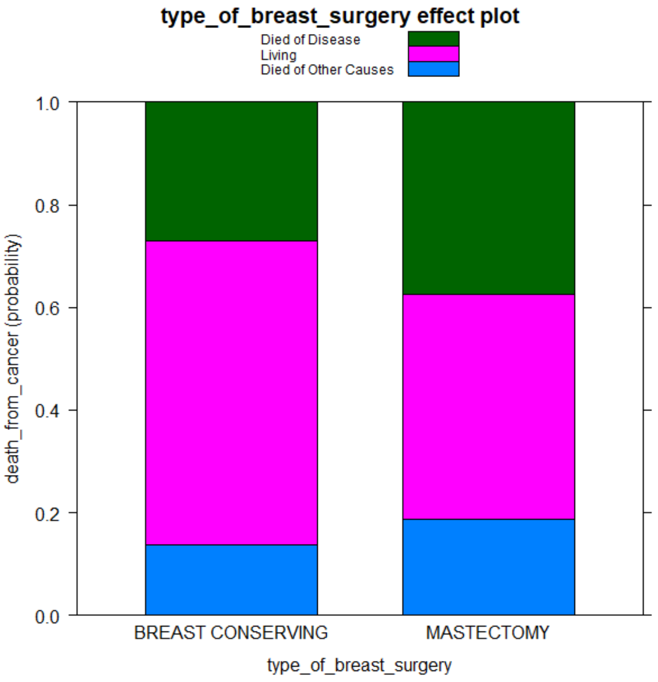


During the exploration of the genomic variables, CCCTC-Binding Factor (*ctcf*) is a transcriptional regulator protein gene that was not correlated with any other gene variable but was significant because this gene is directly associated with invasive breast cancer. The genomic variables could be grouped into three distinct factors: genes associated with the Cyclin Dependent Kinase (CDK) family, genes associated with tumor suppressors, and genes that are associated with cell growth that may or may not be related to gene mutation. The genomic variables were not as salient in predicting overall survival years or attributed to the actual cause of death from cancer. Instead, the genomic variables allowed for key domain knowledge in understanding cancer progression in patients. Ultimately, the clinical variables, specifically the clinical indexes taken during medical examinations, stage of tumor, and type of treatment was more salient in the outcome of the cancer.

Regularized regression was performed to predict the overall survival years of a cancer patient, which although resulted in a lower-than-expected predictive model, still improved the overall performance of the model greatly, and did show some correlation between the variables and the parameter of interest. The model does show that clinical indexes, which is determined during medical examinations and is classified based on various clinical measurements, in addition to the tumor size and stage of the cancer, has some relation to the patient's overall survival. Key patient attributes such as general health and well-being of a patient and length of time of patient’s treatment plans may better predict the parameter of interest. The final model shows that additional patient clinical data would be needed to explore and produce a more adequate model for predicting survival years in cancer patients.

Multinomial logistic regression was performed to predict the categorical placement of death from cancer to the independent variables which showed that the clinical variables played a larger role in determining the outcome of death from cancer than the genomic variables. The three potential outcomes for this analysis were death from cancer, death from other causes, or still living and the value of each variable changes, the probability of each outcome also changes. The final model had an accuracy of 65% and showed that the Nottingham Prognostic Index and the type of breast surgery are more related to death from cancer. The Nottingham Prognostic Index encompasses the size of the tumor, the number of involved lymph nodes, and the grade of the tumor. The relationship of this variable with the probability of death from cancer can be showcased in a line graph. As we can see, the higher the index, the higher the probability of death from cancer. The stacked graph of the type of breast surgery and probability of each outcome shows that for breast conserving breast surgery, the probability of living is very high, while the mastectomy surgery had a lower probability for living.





Given that this analysis is an initial and exploratory analysis, the consensus shows that key clinical indexes which are measures of patient clinical attributes are more significant in modeling survival in breast cancer patients and shown to be more salient in determining the cause of death as cancer. Ultimately, this analysis provided a better understanding of the relationship between clinical variables and genomic variables and how they influence the outcome of cancer patients such as overall survival years and death from cancer. While the models created were not particularly effective at prediction and may not be effective for practical use, the results showed that the clinical variables provided greater insight into the parameter of interests than the genomic variables. Increasing the number of clinical variables, specifically those that gage a patient's overall health, would be beneficial in modeling a patient's long-term survival. Key genomic variables were identified from this analysis that are directly associated with tumor suppressors and growth regulators, which may have influence on cancer outcomes. As gene clustering and bioinformatics is a specific domain, further research could be performed in gene editing with techniques like CRIPSR, which may improve further analysis.

Technical Analysis and Report

Data Preparation and Preprocessing

The breast cancer gene expression dataset was acquired through the METABRIC database and was downloaded from cBioPortal. The dataset comprises a total of 1,904 entries and 693 variables. Each entry represents a breast cancer patient, the patient's clinical and medical attributes, and the patient's gene expression attributes. There are a total of 663 genomic variables and 30 clinical variables in the original dataset. The original dataset had missing values, missing descriptions, and blank entries in many of the clinical variables. These entries were omitted from the analysis. Likewise, 142 of the genomic variables had missing or zero values, these columns were also omitted from the analysis, reducing the number of genomic variables to 488 at the start of the analysis.

Clinical Variables

The first variable in the clinical variables, patient_id, an identification marker was removed as it holds no unique value. Of the remaining 29 variables representing columns [1: 31], 8 of the variables are numeric, 4 of the variables are binary, 16 of the variables are categorical, and 1 variable is ordinal. Using domain knowledge, all 8 of the numeric variables were selected for the analysis since these variables represented attributes that could easily be obtained through medical examinations including age of cancer diagnosis and size of tumor from imaging technologies. Three binary variables were included in the analysis representing if a patient had a type of treatment for the cancer, which was used as a numeric variable for the analysis. Three categorical variables were selected: *type_of_breast_surgery*, *cancer_type_detail*, and *death_from_cancer*. These variables were selected to explore if the type of surgery impacted survival since a surgery type may or may not be available to a patient or a patient may not choose surgery as a treatment option, and to see if the specific type of cancer attributed to the death or survival of the patient. Lastly, the ordinal variable, tumor_stage, a significant variable since the stage of cancer for a patient determines their treatment options and affects their chances of survival is included in the analysis.

The variable *overall_survival_months* was transformed from months to years, to maintain consistency with *age_at_diagnosis*, which is also measured in years. Keeping the same metric allows for correlations to be better observed in the data. The variable was renamed as *overall_survival_years* and was selected as the main parameter of interest for this analysis. This response variable was selected because of its practical application to see if there is correlation between clinical and genomic variables and its effect on survival rate or length of survival after cancer diagnosis. The goal of the analysis is to determine if cancer treatments such as therapies and surgeries, tumor and lymph node indexes measured during medical examinations, or genetic mutations of a patient affect their survival and prolonged living after prognosis and intervention. An additional parameter of interest, *death_from_cancer* was selected after determining that a multinomial logistic regression would be insightful using this response variable since the variable has three levels: living, death due to cancer, and death due to other causes. The goal of this analysis is to determine if the clinical and genomic variables affect the chances of survival outcome and if that outcome is attributed to the actual cancer or due to other causes.

In cleaning the clinical variables, *mutation_count* had significant entries of missing values or non-applicable variables. These entries were omitted and removed from the dataset. Data cleaning also had to be performed on the 3 categorical variables: t*ype_of_breast_surgery*, *cancer_type_detail*, and *death_from_cancer*. For all three categorical variables, all blank entries were omitted. Type of Breast Surgery has two levels: breast conserving (only the part of the breast that contains the cancer is removed) and mastectomy (remove all of the breast tissue). This variable was transformed as a binary variable and for the analysis will be treated as numeric. Cancer Type Detail originally had 5 levels: Breast Invasive Ductal Carcinoma, Breast Mixed Ductal and Lobular Carcinoma, Breast Invasive Lobular Carcinoma, Breast Invasive

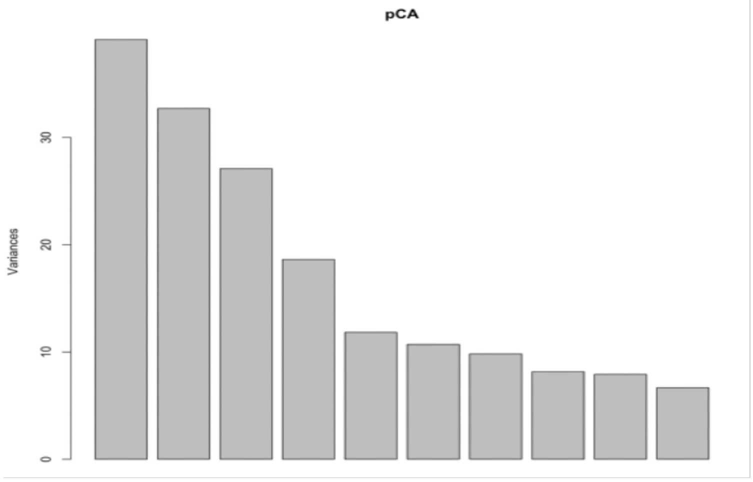
Mixed Mucinous Carcinoma, and Metaplastic Breast Cancer. However, very few entries had Metaplastic Breast Cancer as a cancer type and entries that had this type were removed after cleaning the *mutation_count* variable. This resulted in *cancer_type_detail* having 4 levels with any entry with missing description or missing text in the entry, omitted from the analysis. All of the selected clinical variables and their descriptions can be viewed in the *Appendix A1*.

Looking at the distribution of the clinical variables and the correlations with the parameter of interest, overall survival years, a mostly normal distribution can be seen with several variables including *overall_survival_years*, *age_at_diagnosis*, *nottingham_prognostic_index*, and tumor_stage (*see Appendix A2*). Tumor stage is most correlated with the overall survival years. This ordinal variable affects a patient differently depending on the stage of cancer and due to the correlation plots showing a high correlation to the parameter of interest, it was determined that tumor stage would be a good variable to use as an interaction term in the regression models. Variables that have a very skewed distribution included the nodes examined prognostic index, which is skewed because this index is measured based on patients who have had surgery as part of their treatment in which some patients may not have gotten surgery as treatment, mutation_count which is skewed to left, most likely due to the high number of missing values for this variable but also a higher number of mutations at a certain stage than at other stages may have led to this, and lastly tumor size which is similar to mutation count in skewed distribution. Due to the nature of the dataset being clinical patient data, none of the variables will be further normalized. Instead, tumor stage, tumor size, and mutation count will be analyzed as interaction terms during the model building to see how they improve the model.

Genomic Variables

The 663 genomic variables are numeric and represent m-RNA levels z-score 331 genes and 175 gene mutations. Each m-RNA z-score represents the relative expression of an individual gene and tumor to the gene's expression distribution in a reference population. The reference population is all samples in the study. The returned value indicates the number of standard deviations away from the mean of expression in the reference population. This measure is useful to determine whether a gene is up or down regulated relative to the normal samples of all other tumor samples (CBioPortal). Since all of the genomic variables have been calculated through a z-score, all of the variables exhibit normal distribution as they have already been z-score normalized.

Genomic variables in columns [521: 663] had all zero values or missing values. This most likely represents non-mutated genes or missing values from sampled studies. Since this analysis will focus on mutated genes, columns [521: 663] have been removed from the analysis. The remaining 488 genomic variables will be reduced using dimensional reduction through principal component analysis. Due to the large number of variables and since all of these variables have been z-scored, using principal component analysis to reduce the dimensionality allows for the capture of 70-90% of the variance within the four to six components or less. In this initial reduction, the goal is to reduce the number of variables as much as possible and evaluate the scree plot to determine the number of components. Performing a scaled principal component analysis, resulted in the following scree plot:



The scree plot shows that the majority of the variance is captured within the first four components. Looking at the variance of 10 as a marker, starting at the fifth component, we see an evening out of the components, this can be argued as the knee of the plot. Moreover, from this we will select variables that contribute to the variance from the first four components as our reduced variables. Due to the large number of variables, principal factor analysis at a factor of five was performed for the dimensionality reduction which will reduce the variables to keep the key components and remove the remaining variables as unexplainable noise.

The principal factor analysis above shows that at PC4, the cumulative proportion is 0.91 and the cumulative variance is 0.24. For the dimensionality deduction, all variables with rotated loadings of 0.6 or higher in each component, PC1, PC2, PC3, and PC4 were retained. PC5 did not have any variables with a loading above 0.6. All other variables were removed from the analysis. After the dimensionality reduction, the genomic variables have been reduced from 488 to 25. The 25 selected genomic variables are all protein coding genes with several of the genes associated with the regulation of Cyclin-Dependent Kinase (CDK) kinases and other genes associated with growth and cell differentiation. Protein coding sequences account for only a small percentage, less than 2 percent of the genome but these sequences are critical in the production of all human proteins. The selected genomic variables and their descriptions can be viewed in Appendix *A3*.

After cleaning the data, the cleaned dataset contains 1,283 observations (rows), and 40 variables (columns). There are a total of 15 clinical variables: 12 are numeric, 2 are categorical, and one is ordinal. There are a total of 25 genomic variables which are all numeric and selected from the dimensionality reduction. The main parameter of interest for the analysis is overall survival years, which will predict the survival time of cancer patients based on the clinical and genomic variables. The secondary parameter of interest for the analysis is death by cancer, which will predict the probability of placement into one of three levels with death from other causes used as the reference level that living and death from cancer will be compared to in determining if clinical or genomic attributes more likely influences the cause of death directly due to cancer.

Exploration of Clinical Variables

To explore the data, an analysis of the 15 clinical variables selected through the data preparation from the breast cancer dataset from the METABRIC database will be explored. For this exploratory analysis, 3 binary variables were converted to numeric such as *chemotherapy*, *hormone therapy* and *radio therapy* indicating whether a person is taking therapy as treatment or not. Summation of all numeric data provided 12 variables to kick start the analysis. For the 2 categorical variables, *cancer type detailed* and *death from cancer*, both these variables are explored using a contingency table. In addition, one final numeric variable, *tumor stage,* was determined to be an ordinal factor, due to its high significance and correlation to all other clinical variables. Polychoric factor analysis has been performed on the ordinal factor to interpret its correlation and relationship with all other clinical variables.

Principal Factor Analysis & Common Factor Analysis

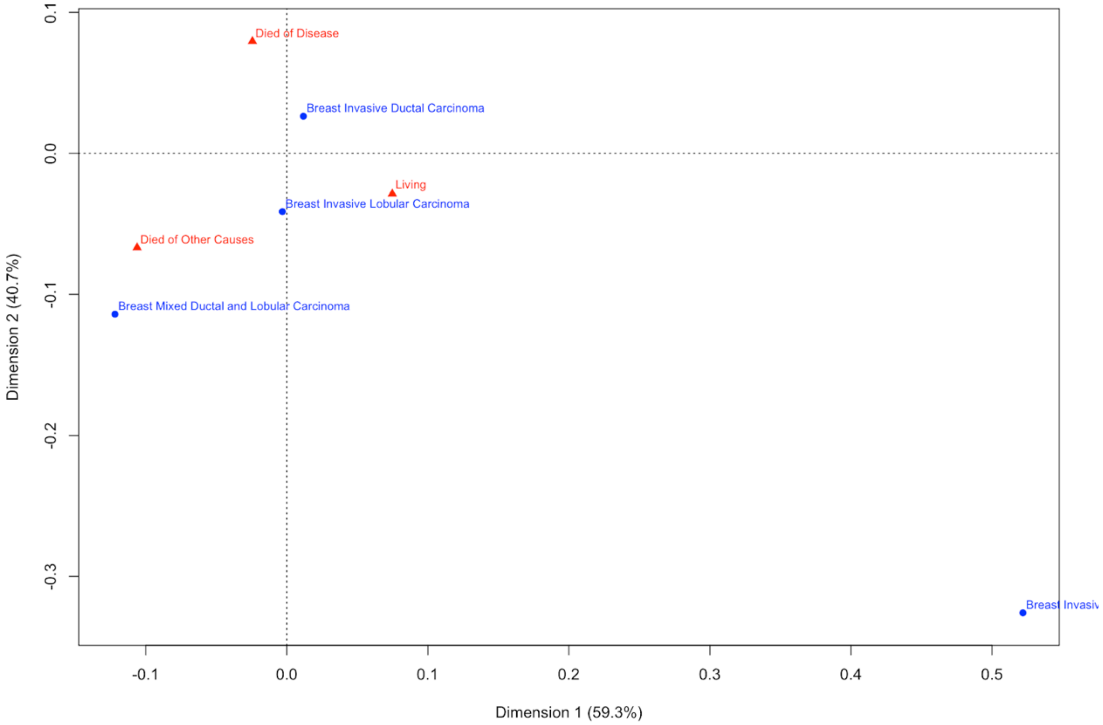
To determine the number of factors to select for the factor analysis, correlation plots were examined, and principal component analysis was performed. We start by visualizing the 12 clinical variables through a correlation plot which transforms the data into a correlation matrix for visualization. Ordering the correlation by angle of eigenvector and through the ellipse method using statistical software R, displayed 2 clusters of degree of multicollinearity between the variables around the data. One variable, Nottingham Prognostic Index showed high collinearity around the data (*see Appendix H*). Further diving into factor analysis gave us 4 factors from the elbow method to interpret the data which captured around 61.6% variance in data. In addition, the principal components reveal that the cumulative variance captured at PC4 is 61.6%, with PC3 capturing 49.1% and PC2 capturing 36.3%. The majority of the variance is captured in PC1, which captured

22.9% of the variance. Thereafter, PC2 captured around 12.8% of the variance, PC3 captured around 12.5% of the variance, and PC4 only captured around 12.5% of the variance. Subsequent components captured even less variance, which confirms that four components are sufficient for factor analysis (see *Appendix B1*) whereas parallel analysis suggests 5 factors (see *Appendix B2*). The loadings revealed 4 distinct groups where overall survival years is negatively correlated. Followed by common factor analysis gives us an understanding of the correlations between the variables. Thus, analysis has been preceded by taking 4 factors into consideration for common factor analysis (see *Appendix B3*) as well which showed that 4 factors were better than 5 factors in comparison (*see Appendix B4*).

RC1 represents low survival rates, which the major contribution in this component is nottingham prognostic index which determines the prognosis following breast cancer surgery. Thus, other components such as neoplasm histologic grade, lymph nodes examined positive, and tumor size all have contribution to this component. Chemotherapy is also a significant contribution and lastly survival years is negatively correlated which illustrates that the tumor has been invaded and chemotherapy may not work as treatment. RC2 represents hormone therapy, which the major contribution in this component is age at diagnosis and determines the patient's age at the time of prognosis. Chemotherapy is negatively correlated here which shows that determining on age of the patient, hormone therapy may work better as treatment. RC3 represents cancer characteristics, which both cohort and mutation count are high contributors to this component. These variables are related to gene variables and represents a shared characteristic of relevant gene mutations. Lastly, RC4 represents treatment types, surgery or therapy, which type of breast surgery is negatively correlated in this component and radio therapy has high positive correlation with others.

Correspondence Analysis

Continuing with the categorical variables, correlation has been considered using a contingency table (see *Appendix C1).* Through the exploration, we found that the most commonly found cancer type is "*Breast Invasive Ductal Carcinoma"*. Also, it is captivating that the rarest cancer type is "*Breast Invasive Mixed Mucinous Carcinoma*" which has the highest survival rate of 73%. Evaluating the mosaic plot, showing the 2D representation of the variable's *cancer type detailed* and *death from cancer* confirms the relationship between the cancer type and survival outcomes.



Polychoric Correlation Analysis

For the ordinal factor analysis, tumor stage is confirmed to be the ordinal variable to explore its correlation with all other clinical variables. In the next step, we conducted polychoric correlation to check its correlation (*see Appendix C2*). During the analysis, binning has been performed for appropriateness where tsize refers to tumor size, progindex refers to nottingham prognostic index and agefactor refers to age at diagnosis. Factor analysis considering 4 factors has been performed to check if any other correlations among the data are possible or not. In the analysis, we found that almost 90% of the variance is captured with 4 components and is sufficient for analysis. Tumor stage is correlated with almost all other variables in the first component and shows significant effect on all other variables. Age at diagnosis is captured only in the second component and it shows correlation with the therapies (chemotherapy, hormone therapy) used for treatment of breast cancer. As both therapies are negatively correlated with each other, it depends on the age and the other corresponding factors to determine treatment plans with therapies. Type of breast surgery is captured only in the third component along with radio therapy which is a similar result discovered in the factor analysis. Lastly, tumor size is significant with tumor stage and is captured in the fourth component which is interesting since we did not obtain any information between their correlation from earlier methods. The interpretation was obtained from performing a polychoric factor analysis (*see Appendix C3*) and indicates that 89.4% of the variance around the data was captured within four factors.

Exploration of the clinical variables gave us a better understanding of the relationship of the variables to the parameters of interest and resulted in key findings through the factor, correspondence, and polychoric correlation analysis. There is high multicollinearity around Nottingham Prognostic Index which was expected since this index is measured from other numeric clinical variables. Factor analysis had a good fit with four factors representing low survival rates, hormone therapy, cancer characteristics, and surgery as treatment versus therapy as treatment. During categorical analysis, the common type of cancer and highest rate of survival with another type of cancer were found which were fascinating results. Polychoric correlation analysis with the most significant factor tumor stage gave the best interpretation of all the other variables. All different therapies functional for treatment of breast cancer could be analyzed further when other contributing factors towards this life-threatening disease were provided. Fundamental risk factor is genetic mutations. Yet, many other contributing factors include the spread of tumor cells, physical activity, alcohol consumption status, pregnancy history and stress levels which when imparted in the data would have complemented our analysis unearthing the overall survival rate of women with breast cancer.

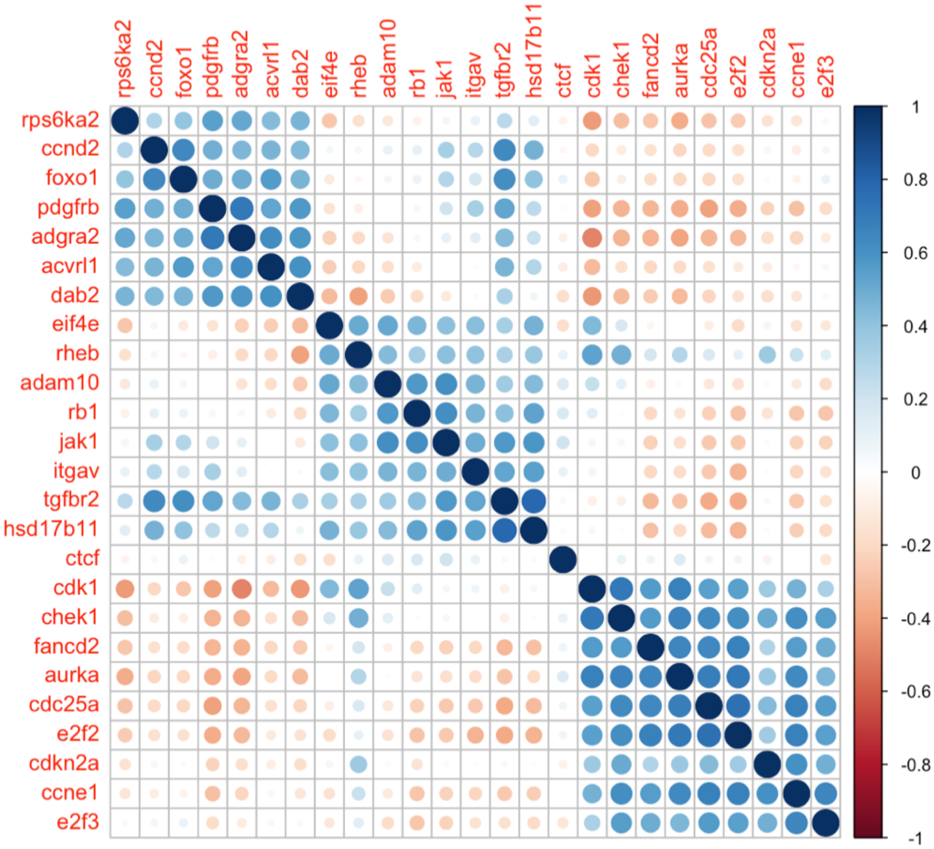
Exploration of Genomic Variables

This analysis explores the 25 genomic variables selected through the dimensionality reduction from principal component analysis on the gene expression variables in the breast cancer gene expression dataset from the METABRIC database. The selected variables are all protein coding genes which account for only 1.5% of the genome but play a significant role as they ultimately lead to the production of all human proteins. To explore the relationship between the variables, this analysis will perform principal factor analysis, common factor analysis, and cluster analysis. The goal of the exploration is to determine significant key genomic variables to be used in modeling regression and to determine if the genes have any meaning or patterns in their characteristics. In this analysis, multidimensionality scaling will be discussed to explore the fit of the data for cluster analysis. In addition, methods for the factor analysis and visualizations will be discussed to show how the factors were determined, significance of the factors, and interpretation of the loadings. The variables will be shown as a good fit for factor analysis and provide significant insight for the parameter of interest.

Principal Factor Analysis & Common Factor Analysis

To determine the number of factors to select for the factor analysis, correlation plots were examined, and principal component analysis was performed. We start by visualizing the 25 genomic variables through a correlation plot which transforms the data into a correlation matrix for visualization. Ordering the correlation by angle of eigenvector and through the ellipse method using statistical software R, displayed a high degree of multicollinearity between the variables and showed three distinct factors. One variable, *ctcf*, was not highly correlated with any other variable, whereas rheb is correlated with almost all other variables (*see Appendix D1*). Ordering the correlation by hierarchical clustering confirmed the existence of three distinct groupings and ctcf as the variable that showed little or no correlation to all other variables. The correlation plots reveal that three or four factors exist, which we will explore further by analyzing the principal components.

Correlation Plot:
All 25 variables,
four distinct groupings,
and ctcf not correlated.



Principal component analysis was performed on the 25 genomic variables using the correlation matrix. The scree plot for the principal components revealed a knee occurring at PC4. Using the variance equals 1 criterion, the first four components are above 1 (*see Appendix D2*). Moreover, the knee and the variance equals 1 both confirm four factors. The principal components reveal that the cumulative variance captured at PC4 is 67.9%, with PC3 capturing 62.7% and PC5 capturing 71.6%. The majority of the variance is captured in PC1, which captured 30% of the variance. Thereafter, PC2 captured around 19% of the variance, PC3 captured around 14% of the variance, and PC4 only captured around 5% of the variance. Subsequent components captured even less variance, which confirms that four or three components is sufficient for factor analysis (*see Appendix D3*).

An initial principal factor analysis using four factors and varimax rotation was performed. The loadings revealed three distinct groupings with ctcf separated into its own factor, RC4. RC1 included *cdk1, ccne1, cdc25a, cdkn2a, e2f2, e2f3, qurka, chek1,* and *fancd2*. RC2 included *rb1, jak1, adam10, eif4e, itgav, reheb,* *tgfbr2*, and *hsd17b11*. RC3 included *ccnd2, acvrl1, dab2, foxo1, pdgfrb, rps6ka2, adgra2*, and tgfbr2 which was also included in RC2, and *cdk1* in negative relationship which was included in RC1 (*see Appendix D4*). Common factor analysis was performed with four components as a comparison and to see if the rotated components would separate more clearly. The results showed that common factor analysis did not perform any better than principal factor analysis at four factors. The variables for each factor remained the same, with the fourth factor including *ctcf* and *eif4e* in negative coefficients (*see Appendix D5*). Since *ctcf* becomes its own factor from the factor analysis, we run the factor analysis again but using three factors as a comparison. The results show that with three factors for both principal factor analysis and common factor analysis, *ctcf* falls out completely and is not included in any of the factors. In addition, the variables are grouped more clearly with three factors, with *cdk1* separating to the first factor and only *tgfbr2* not separating entirely and grouped into one of two factors (see Appendix D6).

The results from the initial factor analysis confirms that *ctcf* is not correlated with any other variable and becomes its own factor, confirming the initial visualization from the correlation plot. To validate the goodness of fit of the variables and if factor analysis is appropriate for the data, the Kaiser-Meyer-Olkin (KMO) factor adequacy was performed with an overall MSA (measure of sample adequacy) of 0.9, which is strong and shows a high factor stability. All variables had an MSA of over 0.8, except ctcf which had an MSA of 0.47, confirming that this variable may not be suitable with other factors (*see Appendix D7*). To interpret the factors and components we will perform the factor analysis again excluding *ctcf*.

Performing the principal factor analysis with three factors on the 24 genomic variables, resulted in the same factor rotations as the initial principal factor analysis (*see Appendix D8*). Common factor analysis with three factors on the 24 genomic variables performed better than principal factor analysis in separating the variables into distinct factors, but the total cumulative variance is slightly lower at around 60%, compared to 65% with principal factor analysis. In both factor analyses, each factor captured around 20% of the variance, with the first factor capturing a slightly higher percentage.

The first factor (RC1) appears to represent genes that are cyclin-dependent kinase, proteins in the kinase family, or proteins associated with regulating the cyclin kinase family (*see Appendix D9*). Cyclin Dependent Kinase 1(cdk1) is directly associated with breast cancer. The loss of Cyclin Dependent Kinase Inhibitor 2A(cdkn2a) is directly associated with the development of many cancers. The majority of the genes in this factor are associated with retinoblastoma (eye cancer), such as E2F Transcription Factor 2 (e2f2) which is critical in the control of cell cycle and the action of tumor suppressor proteins. Cancers including breast cancer cause dysregulated CDK/cyclins, which will cause instability in the coordinated cycle of cell growth and proliferation, contributing to the uncontrolled proliferation of cancer cells (Peyressatre 2015). Moreover, the first factor represents genes that have been documented to cause cancers or when unregulated results in mutations causing tumors and cancers.

The next factor (RC3) appears to be receptor, encoder, or adaptor proteins that serve the serine-threonine kinase domain or subdomain (*see Appendix D10).* This domain is associated with the CDK/cyclin family. Majority of the genes in this factor are associated with cell growth or control of cell growth and division. Cyclin D2 (*ccnd2*) is highly associated with cancers if unregulated. Certain genes in this factor such as Forkhead Box O1(*foxo1*) do not have known functions yet. This factor includes many genes that encode proteins for other genes and a loss or non-regulation of these genes may result in cell differentiation. This factor includes a negative relation with *cdk1*, which shows that perhaps this factor may be assisting with suppressing cancer, since cdk1 directly causes cancer.

The last factor (RC2) appears to represent genes that are either tumor suppressors, binding proteins, and proteins that can stabilize or regulate cancer progression (*see Appendix D11*). RB Transcriptional Corepressor 1(*rb1*) is the first known tumor suppressor gene. Janus Kinase 1(*jak1*) and Integrin Subunit Alphia V (itgav) both are genes that can regulate or restrict cancer progression. Moreover, this factor represents genes that have been documented to suppress or regulate the development and progression of tumors. Lastly, CCCTC-Binding Factor (*ctcf*) is a transcriptional regulator protein associated with 11 highly conserved zinc finger (ZF) domains. Mutations of this gene have been directly associated with invasive breast cancers. As this gene is in a different domain than most of the other genes, this explains why the gene is not correlated with other genes.

Multidimensionality Scaling & Cluster Analysis

To further explore the relationship between the genomic variables, an exploratory cluster analysis was performed. First, multidimensional scaling was performed on the genomic variables to determine the goodness of fit for cluster analysis. The stress resulted in a value of 0.187, which is much greater than 0.1 but not greater than 0.2, in other words the genomic variables are most likely not a good fit for cluster analysis (*see Appendix E1*). When looking at the Shepherd's Diagram, the distances were wide with no clear step line, further confirming that the data may not be good for cluster analysis (see *Appendix E*2). The plot of the multidimensional scaling on two dimensions did not show any clear distinct clusters, but instead, showed one dense cluster with possible outliers (*see Appendix E3*).

As an exploration, the following cluster analysis approaches were performed on the data: density, k-means, k-medoids, and hierarchical. Density clustering failed and resulted in zero clusters. K-medoids, which clusters based on the most central object and hierarchical clustering, which clusters based on subclusters performed the best. To determine the optimal number of clusters, the average silhouette method and the gap statistic method was used. The average silhouette method measures the quality of the clustering and determines how well each object lies within its cluster. Using the average silhouette method for the k-medoids approach resulted in k = 2, which is low, since a high average silhouette width indicates good clustering (*see Appendix E4*). The gap statistic method compares the total intracluster variation for different values of k with their expected values under a distribution with no obvious clustering. Using the gap statistic method for the hierarchical approach resulted in k = 2, which is the same results from the silhouette method (*see Appendix E5*).

For the hierarchical clustering, first the dissimilarity values were obtained using the similarity matrix on the data and those values were used to plot the hierarchical cluster and obtain the dendrogram (*see Appendix E6*). The dendrogram has a tall height with many groupings and even at k = 2, the groupings are not distinct. The resulting cluster confirms that no distinct clusters were found with the first cluster overlapping with the majority of the second cluster (*see Appendix E7*). When increasing to k = 3, one cluster overlaps with the other two clusters and the cluster plot does visually appear more separated between clusters than with two clusters, but the first cluster still overlaps heavily with the other two clusters resulting in difficult interpretation and no clear distinction. K-medoids cluster analysis using k = 2 performed slightly better although the resulting clusters are very similar to the results from hierarchical clustering with the first cluster having slightly less overlap to the second cluster (*see Appendix E8*). With the results of these cluster plots and with the multidimensional scaling stress test showing a value close to 0.2, we can conclude that distinct clusters could not be found in the data. K-medoids clustering resulted in the best cluster plots at k = 2 but heavy overlap between the clusters exists and clear distinctive clusters could not be determined.

The exploration of the genomic variables resulted in several key findings from factor analysis and cluster analysis. The genomic variables could not be clustered into any distinct clusters and that clustering was not a good fit for the variables. Gene clustering is a specific domain and more advanced methods for exploring the genes could be applied than the clustering techniques used in this exploration. For factor analysis, three distinct factors were determined with the first factor representing genes associated with CDK/cyclins and genes regulating or controlling proteins associated with the kinase family, which when left unregulated has been shown to contribute to the proliferation of cancer. Moreover, the first factor can be seen as genes associated with the development of breast cancer by way of the control or regulation of CDK/cyclins proteins and their affiliates. The second factor represents genes that are tumor suppressors or genes that help stabilize or restrict the progression of cancer. Lastly, the last factor represents genes associated with the serine-threonine kinase domains, which are associated with the kinase family. This factor can be seen as genes that code or regulate growth and division of cells in these domains, which when left unregulated will result in cell differentiation. Whereas the first two factors appear to be opposite of each other, first factor contributing to cancer, and the second factor aiding in the progression of cancer, the third factor is less clear since it involves genes that potentially could be tumor suppressors but also involves genes when unregulated could lead to mutations resulting in cancer. Lastly, *ctcf* was not correlated with any factors, but is associated with invasive breast cancer and most likely represents a variable aiding in cancer progression.

Regularized Regression: Overall Survival Years

We attempted to construct a model that would predict survival years of breast cancer patients as this would be valuable information for patients that have been recently diagnosed. For building this model we started with 39 variables including 25 genes and 14 clinical variables. From that starting point the first 4 PCA variables were added from the 25 genes that we had selected out as being important. It is useful to stick to using the PCA values from these genes since another member of the group is going to do common factor analysis so we could interpret our final model if it included any of the PCA variables. The categorical variable "type of breast surgery" was turned into a binary value so that it could be used in the regression. Lastly some interaction variables noted below were added based on some of our previous analysis and hypothesis.

- Radiotherapy and Tumor Size
- Radiotherapy and Tumor Stage
- Chemotherapy and Tumor Size
- Chemotherapy and Tumor Stage
- Hormone therapy and Tumor Size
- Hormone therapy and Tumor Stage

We looked at multiple different regression techniques which are summarized in the table below. There is a noticeable gap between the training set and test sets RMSE for the most basic of the models showing that there is some multicollinearity and overfitting in our data set. Without delving too far into these models they are generally not very good due to their R^2 and the difference in RMSE between training and test sets.

Summary table for different regression models explored:

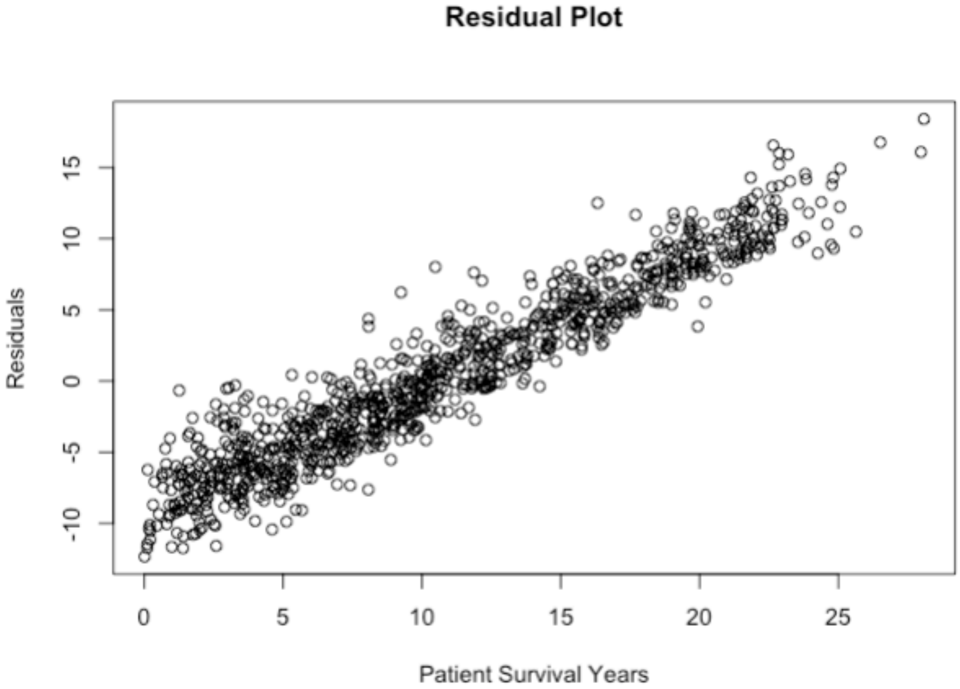
	Base Model	Backwards Step Regression	All Subsets Regression	Ridge Regression	Lasso Regression	Relaxed Lasso Regression
Adjusted R-2	0.188	0.244	0.236	0.237	0.255	0.2652
Training RMSE	5.41	5.58	5.67	N/A	N/A	N/A
Testing RMSE	6.45	6.42	6.36	6.20	6.23	6.24
Parameters	47	12	6	47	13	7

The best performing regression model was the relaxed lasso regression. Using this model, we get a parsimonious model with only 7 variables and the highest adjusted R^2 value along with a lower RMSE on the testing data set than most of the other models. That being said, our model is still not very useful for predicting survival years of breast cancer patients. A mean squared error of 6.2 years is way too large to be effective at predicting the remaining lifespan of a patient. There are clearly some important variables that we are not able to consider with our data set. Our data set lacks a lot of underlying variables that impact the general health of a patient. Obesity, fitness level, diet, blood pressure, heart rate and other medical conditions are just a few of the many variables that we hypothesize could have an impact on the progression of cancer and consequently patient survival after being diagnosed.

Final Model

Survival Years = 13.84 – 0.16 Lymph Nodes Examined Positive – 0.28 Nottingham Prognostic Index – 0.004 Tumor Size -1.10 Tumor Stage - 0.01 chek1- 0.614 rheb – 0.17 PC4

One of the clearest signs of bias in our model is shown in the residual plot below. In a useful model we would expect the residuals to be scattered homoscedasticity around 0, horizontal to the x-axis. Instead, what we see is that our model generally gives predictions near the average of the data set only slightly increasing or decreasing its prediction based on the independent variables. Our model performs poorly at predicting patients' survival years if they differ greatly from the mean.



A more practical use of our model is to look at the variables that ended up being included in our model and compare them to the factor analysis previously performed. Nottingham Prognostic Index and Lymph Nodes Examined positive are the two largest factors of the first principle loading in the clinical data. This gives us confidence in the stability of the model. Tumor size is also in this first factor. It is interesting that PC4 is included in this model over some of the larger principal components. One reason might be that the rheb gene is one of the largest contributors to PC2 rendering it less useful. Also, chek1 is a very large contributor to PC1, PC2 and PC3. So, since the model already selected these genes the first three principal components could be selected out by Lasso due to their multicollinearity.

Multinomial Logistic Regression: Death by Cancer

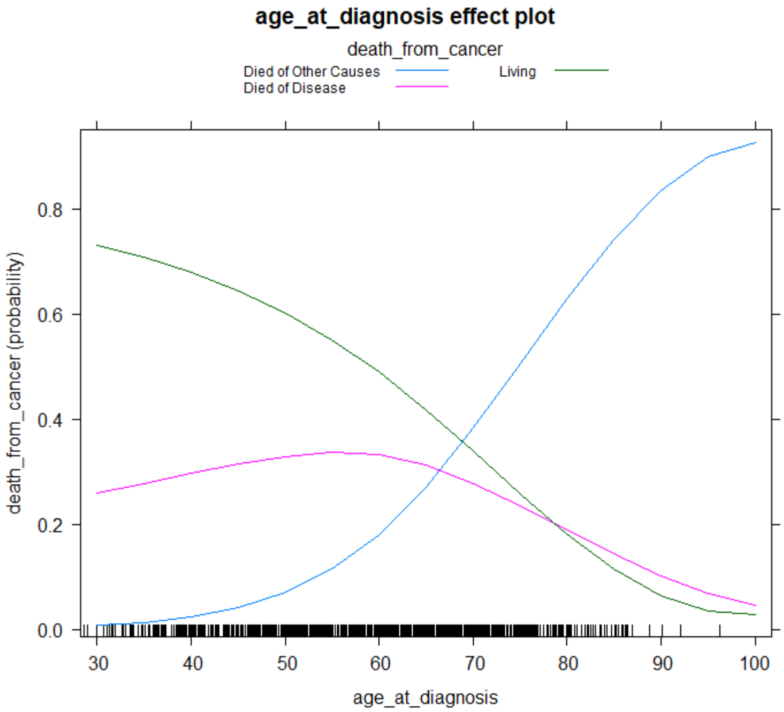
Multinomial logistic regression is a form of regression analysis where the dependent variable is nominal and has more than two categories. This type of regression can have nominal or continuous independent variables, which makes it well suited for the data target data set since some of the variables are categorical. With multinomial logistic regression there are some important considerations, such as needing a large sample size and checking for empty or small cells. Multinomial logistic regression was used instead of collapsing the number of categories into two and conducting logistic regression because it would suffer from information loss. In this analysis, we explored deaths from cancer as the dependent variable, which has three possible outcomes of "died of disease," "died of other causes," and "living."

To begin, an all-variable model was created to initially explore the data and be used as a base to compare subsequent models to. In order to create the multinomial logistic model, the package *nnet* was used for the function *multinom* in R. The variable death_from_cancer was re-leveled so "died of other causes" was the reference level. When applied to the training data set, a model of 71.38% accuracy was created. When it was reapplied to the test data set, the model had an accuracy of 69.9%, indicating that it may not have suffered from overfitting due to the low drop in accuracy. This model was also analyzed to assess the most significant variables at play so we could reduce the number of variables in the model to make it more parsimonious and easier to interpret. From there, we created and analyzed multiple models to understand how well these variables work in predicting the outcome of death from cancer.

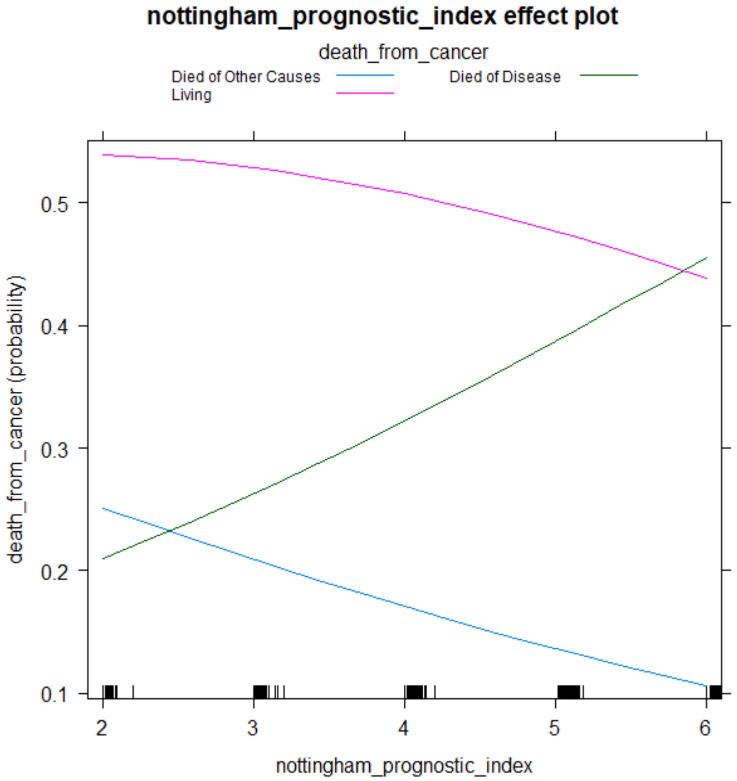
Summary of the Top Three Models and their Accuracies:

Top Multinomial Logistic Regression Model	*Accuracy from Training*	*Accuracy from Test*
All Clinical and Genomic Variables	71.38%	69.90%
survival_years, age_at_diagnosis, type_of_breast_surgery, nottingham_prognostic_index, e2f2, and aurka	66.92%	64.86%
nottingham_prognostic_index, age_at_diagnosis, cohort, mutation_count, type_of_breast_surgery, ccne1, cdc25a, chek1, acvrl1, foxo1, and jak1	59.29%	55.56%

Of the top three performing models, the model with the fewest variables was selected for visualization because of its interpretability and accuracy. It included the variables survival_years, age_at_diagnosis, type_of_breast_surgery, nottingham_prognostic_index, *e2f2*, and *aurka*. The accuracy on the test dataset was 64.86%, which was only a 2% drop from the accuracy of the model on the data used in training. From here, we can interpret the probability of dying from cancer, dying from other causes, and living based on the values of the different variables in the model. The clinical variables played the largest role in all of the models. Clinical variables such as the type of surgery, age, and the Nottingham Prognostic Index play a significantly bigger role in determining the outcome of dying from cancer than the gene expression variables.



With an increase in age, the probability of dying from cancer decreases, but so does the probability of living. This is because the probability of dying from other causes increases dramatically. At the age of 30, the probability of dying from cancer is relatively low and slowly increases until about the age of 60, where it starts to drop again due to drastic increase in death from other causes.



Another interesting trend is with the Nottingham Prognostic Index, where we found that the higher the index number, the higher the probability of death from disease. Death from other causes decreases with the higher index due to the higher likelihood of dying from cancer. The outcome of living also goes down slightly with the higher index number. For example, a Nottingham Prognostic Index of 2 has a 55% probability of living and 20% probability of dying from cancer, but an Index of 6 has a roughly 43% probability of both living and dying from cancer.

Conclusion

Exploration of the breast cancer gene expression dataset through the METABRIC database revealed key findings in relation to clinical and genomic variables and their significance in predicting the overall survival years in breast cancer patients. In exploring the clinical variables, it was determined that tumor stage was the most significant variable through polychoric correlation and linear discriminant analysis. Understanding the stage of the tumor allowed for a better understanding of all other clinical variables as the stage of the cancer directly dictated the results of all other clinical variables. The clinical variables were a good fit for factor analysis with four factors, representing low survival rate, hormone therapy, cancer mutation characteristics, and type of treatment. All different treatment therapies aiding in the progress of a patient's treatment depending on their age of prognosis and the cancer type had impact on the outcome of the disease. The genomic variables were also a good fit for factor analysis with three factors representing genes associated with the CDK/Kinase family that when left unregulated results in the proliferation of cancer cells, genes associated with tumor suppressors, or genes that are growth regulators that may or may not be associated with cancer progression. One gene, ctcf, is not correlated with any other gene but is directly associated with invasive breast cancer and is significant in the model.

From the factor analysis, tumor stage and *ctcf* were shown to be unique and significant in different ways. Both variables are significant in the final regression model for the overall survival years. Almost all models that were built to predict overall survival years supported the clinical and genomic data from the factor analysis. The final regression model is a relaxed LASSO model, with an adjusted R-square of 0.2652 and included features: Lymph Nodes Examined Positive, Nottingham Prognostic Index, Tumor Size, and Tumor Stage, *jak1, chek1, rheb, and ctcf*. Although the adjusted R-square is low for practical use, especially since the RMSE was 6.24, the model does show that clinical indexes, which is determined during medical examinations and is classified based on various clinical measurements, in addition to the tumor size and stage of the cancer, has some relation to patient's survival. Nottingham Prognostic Index proved to be significant in multinomial logistic regression for the death from cancer as the higher the index score, the higher the probability of death from cancer. This index along with key clinical variables such as type of surgery and age of prognosis had a larger role in determining the outcome of death from cancer than the genomic variables. This analysis shows that mitigation efforts in the patient's clinical attributes including when the cancer is diagnosed, the stage of the cancer, and treatment options are more salient in determining the outcome of the patient. Moreover, prevention is better than a cure and after prognosis, tumor cells need to be detected early to determine best treatment.

Throughout our analysis the prominence of clinical variables with relation to the well-being of patients was clear. Without additional medical knowledge about the patients, it is difficult to create a model that can accurately predict the well-being of a cancer patient. We think that increasing the number of clinical variables, specifically those that gage a patient's overall health, would be beneficial in modeling a patient's long-term survival. Outside of gathering additional data for further analysis we did identify some key genes in relationship to survival of breast cancer patients. Genes like *cdk1, rheb, ctfc and cdkn2a* have all been identified as possible targets for new gene editing techniques like CRIPSR. An analysis comparing similar patients that do and do not have gene editing performed to fix their genes could lead to a better understanding of which gene mutations are most significant in breast cancer patients. Our analysis just begins to scratch the surface of the work to be done with regards to breast cancer, but we hope that it can be a starting point for further analysis that can lead to a cure.

Work Cited / References

“Breast Cancer Gene Expression Profile (METABRIC)”. cBioPortal for Cancer Genomics. https://www.cbioportal.org/ Accessed 01 April 2021

Peyressatre, Marion, et al. "Targeting cyclin-dependent kinases in human cancers: from small molecules to peptide inhibitors." *Cancers* 7.1 (2015): 179-237. https://www.mdpi.com/2072-6694/7/1/179 Accessed 20 May 2021

Pereira, Bernard, et al. "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes." Nature communications 7.1 (2016): 1-16. https://www.nature.com/articles/ncomms11479

UCLA. (n.d.). “Multinomial Logistic Regression - R Data Analysis Examples”. Institute for Digital Research & Education. https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/. Accessed 20 May 2021.

GeneCards®: The Human Gene Database. GeneCards. https://www.genecards.org/ Accessed 20 May 2021.

Safran, Marilyn, et al. "GeneCards Version 3: the human gene integrator." *Database* 2010 (2010). Accessed 29 May 2021.

“Treatment and Side Effects.” Breastcancer.org, 20 April 2021, https://www.breastcancer.org/treatment. Accessed 01 June 2021.

“Breast Cancer Statistics and Resources.” Breast Cancer Research Foundation, 18 May 2021, https://www.bcrf.org/breast-cancer-statistics-and-resources. Accessed 06 June 2021.

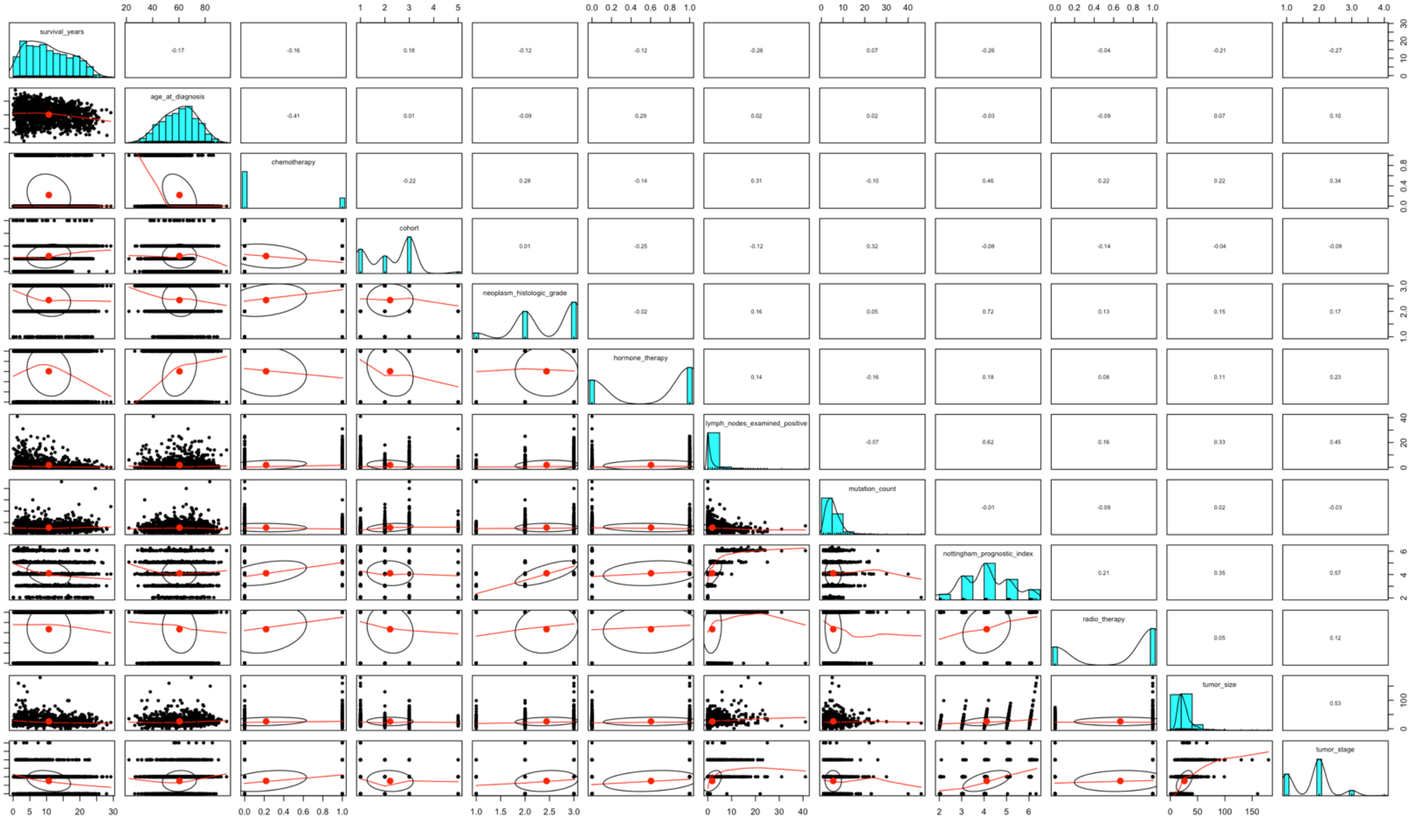
Appendix: Technical Analysis

Appendix A: Data Preparation

Appendix A1: Clinical Variables

Variable	Type	Type for Analysis	Description
*overrall_surival_years (*parameter of interest*)	Numeric	Numeric	Duration from the time of the intervention to death (in years). Transformed to years from months.
age_at_diagnosis	Numeric	Numeric	Age of the patient at diagnosis time (in years)
cohort	Numeric	Numeric	Cohort is a group of subjects who share a defining characteristic (value from 1 to 5)
neoplasm_histologic_grade	Numeric	Numeric	Determined by pathology by looking the nature of the cells, and determining if they look aggressive or not (value from 1 to 3)
lymph_nodes_examined_positive	Numeric	Numeric	Samples of the lymph node during the surgery and see if the lymph nodes evolved by the cancer.
nottingham prognostic index	Numeric	Numeric	Used to determine prognosis following surgery for breast cancer. Value is calculated using three pathological criteria: the size of the tumor; the number of involved lymph nodes; and the grade of the tumor.
mutation_count	Numeric	Numeric	Number of genes that have relevant mutations.
tumor_size	Numeric	Numeric	Tumor size measured by imaging techniques.
tumor_stage	Ordinal	Ordinal	Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread.
chemotherapy	Binary	Numeric	Whether or not the patient had chemotherapy as a treatment (Binary: Yes/No)
hormone_therapy	Binary	Numeric	Whether or not the patient had hormonal as a treatment (Binary: Yes/No)
radiotherapy	Binary	Numeric	Whether or not the patient had radiotherapy as a treatment (Binary: Yes/No)
type_of_breast_surgery	Categorical	Binary/Numeric	Binary: 2 Levels 1 = Breast Conserving (only the part of the breast that contains the cancer is removed). 2 = Mastectomy (remove **all** tissue breast, method of treatment or prevention of cancer)
cancer_type_detail	Categorical	Categorical, 4 Levels	4 Levels: Breast Invasive Ductal Carcinoma, Breast Mixed Ductal and Lobular Carcinoma, Breast Invasive Lobular Carcinoma, Breast Invasive Mixed Mucinous Carcinoma
** death_from_cancer (***parameter of interest*)	Categorical	Categorical, 3 Levels	3 Levels: Still Living, Death from Cancer, Death from Other Causes

Appendix A2: Correlation and Distribution of Clinical Variables



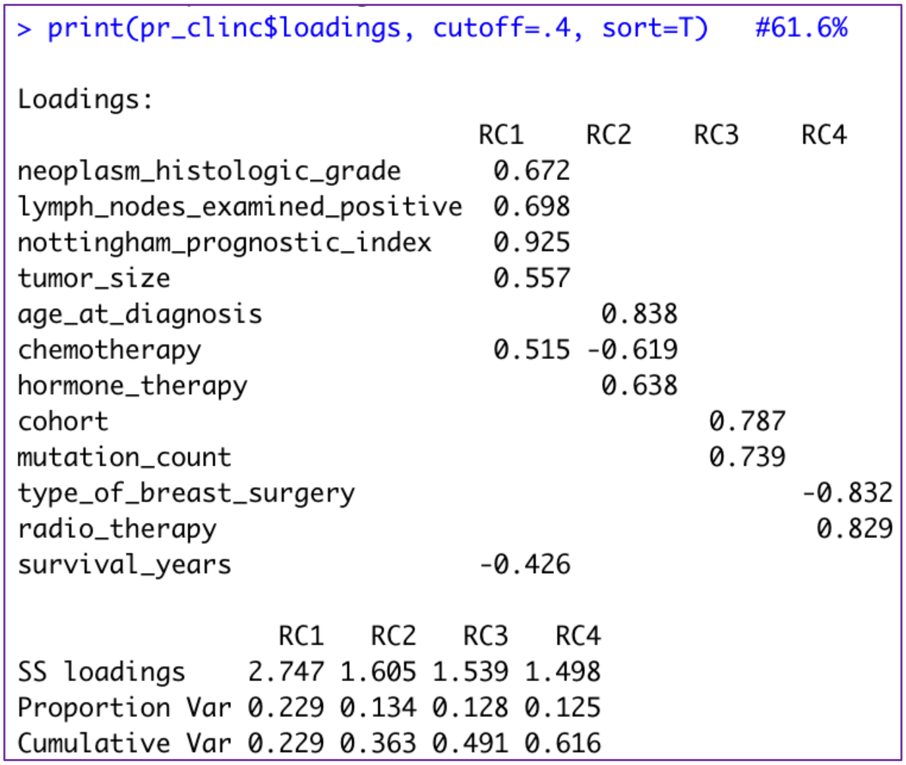
Appendix A3: Genomic Variables

Variable	Description (all protein-coding genes):
rb1	RB Transcriptional Corepressor 1: Protein encoded by this gene is a negative regular of the cell cycle and was first tumor suppressor gene found. Associated with Retinoblastoma and Small Cell Lung Cancer.
cdk1	Cyclin Dependent Kinase 1: Protein encoded by this gene is part of the Ser/Thr protein kinase family. Catalytic subunit of highly conserved kinase complex, essential for G1/S and G2/M phase transition of eukaryotic cell cycle. Associated with Retinoblastoma and Breast Cancer.
ccne1	Cyclin E1: Protein encoded by this gene is part of the highly conserved cyclin family. Cyclins function as regulators of CDK kinases. Overexpression of this gene has been observed in many tumors. Associated with clear cell Adenocarcinoma of the Ovary and Retinoblastoma.
cdc25a	Cell Division Cycle 25A: Part of the CDC25 family of phosphatases. Required for progression from G1 to the S phase of the cell cycle. Activates the cyclin-dependent kinase CDC2. Associated with Retinoblastoma and Ataxia-telangiectasia.
ccnd2	Cyclin D2: Protein coded by this gene belongs to the highly conserved cyclin family. Cyclins function as regulators of CDK kinases. Has been shown in many cancer types if unregulated.
cdkn2a	Cyclin Dependent Kinase Inhibitor 2A: Gene generates several transcript variants which different in their first exons. Gene is frequently mutated or deleted in a wide variety of tumors and is known to be an important tumor suppressor gene. Loss of this gene shown to be significant in many cancers.
e2f2	E2F Transcription Factor 2: Protein encoded by this gene is a part of the E2F family of transcription factors. Plays a crucial role in control of cell cycle and action of tumor suppressor proteins. Associated with Retinoblastoma and Encapsulated Thymoma.
e2f3	E2F Transcription Factor 3: Encodes a member of a small family of transcription factors that function through binding of DP interaction partner proteins. Associated with Retinoblastoma and Bladder Cancer.
jak1	Janus Kinase 1: Encodes a membrane protein that is part of a class of protein-tyrosine kinases (PTK). Gene plays a crucial role in effecting the expression of genes that mediate inflammation, epithelial remodeling, and metastatic cancer progression. Associated with Autoinflammation, Immune Dysregulation, and Eosinophilia.
adam10	ADAM Metallopeptidase Doman 10: Cell surface proteins with a unique structure possessing both potential adhesion and protease domains. Associated with reticulate a Reticulate Pigmentation of Kitamura and Alzheimer Disease.

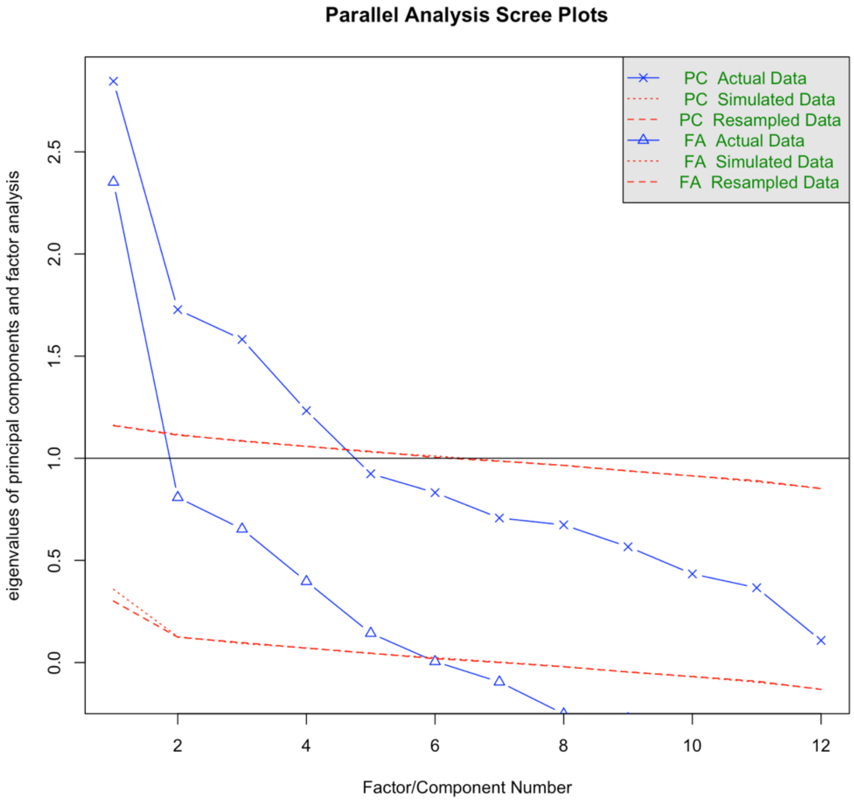
acvrl1	Activin A Receptor Like Type 1: Encodes a type 1 cell-surface receptor for the TGF-beta superfamily of ligands. Shares high degree of similarity. To the serine-threonine kinase subdomains. Associated with Telangiectasia and Hereditary Hemorrhagic.
aurka	Aurora Kinase A: Protein encoded by this gene is a cell cycle-regulated kinase that appears to be involved in microtubule formation and/or stabilization at the spindle pole during chromosome segregation. Gene may play a role in tumor development and progression. Associated with Colorectal Cancer.
chek1	Checkpoint Kinase 1: Protein encoded by this gene is part of the Ser/Th protein kinase family. Required for checkpoint cell cycle arrest in response to DNA damage or the presence of un-replicated DNA. Associated with Ataxia-Telangiectasia and Li-Fraumeni Syndrome.
dab2	DAB Adaptor Protein 2: Encodes a mitogen-responsive phosphoprotein. Expressed in a normal ovarian epithelial cell but is downregulated or absence from ovarian carcinoma cell lines, suggesting its role as a tumor suppressor. Associated with Teratocarcinoma.
eif4e	Eukaryotic Translation Initiation Factor 4E: Protein encoded by this gene is a component of the eukaryotic translation initiation factor 4F complex, which recognizes the 7-methlguanosine cap structure at 5’ end of messenger RNAs. Associated with Autism and Pervasive Development Disorder.
foxo1	Forkhead Box O1: Part of the forkhead family of transcription factors. Specific function has not yet been determined, but it may play a role in myogenic growth and differentiation. Associated with Rhabdomyosarcoma and Glioma.
itgav	Integrin Subunit Alpha V: Product of this gene belongs to the integrin alpha chain family. Integrins are heterodimeric integral membrane proteins and may regulate angiogenesis and cancer progression. Associated with West Nile Virus and Herpes Simplex.
pdgfrb	Platelet Derived Growth Factor Receptor Beta: Protein encoded by this gene is a cell surface tyrosine kinase receptor for the platelet-derived growth factor family. This gene is essential for the normal development of the cardiovascular system and aids in rearrangement of the actin cytoskeleton. Associated with Premature Aging Syndrome and Kosaki Overgrowth Syndrome.
rheb	RAS Homolog, HTORC1 Binding: Gene is a member of a small GTPase superfamily and encodes a lipid-anchored, cell membrane protein with five repeats of the RAS-related GTP-binding region. Associated with Tuberous Sclerosis and Hemimegaloencephaly.
rps6ka2	Ribosomal Protein S6 Kinase A2: Encodes a member of the RSK (ribosomal S6 kinase) family of the serine/threonine kinases. Activity of this protein has been implicated in controlling cell growth and differentiation. Associated with Coffin-Lowry Syndrome and Autism.
tgfbr2	Transforming Growth Factor Beta Receptor 2: The protein encoded by this gene is a transmembrane protein that has a protein kinase domain and forms a heterodimeric complex with TGF-beta receptor type-1, and binds TGF-beta. Mutations of this gene have been associated with Marfan Syndrome, Loeys-Deitz Aortic Aneurism Syndrome, and the develop of various types of tumors. Diseases associated with TGFBR2 include Loeys-Dietz Syndrome, Colorectal Cancer, and Hereditary Nonpolyposis.
adgra2	Adhesion G Protein-Coupled Receptor A2: Part of the adhesion-GPCR family of receptors. Endothelial receptor which functions together with RECK to enable brain endothelial cells to selectively respond to the Wnt7 signals and establish blood-brain barriers.
ctcf	CCCTC-Binding Factor: Part of the BORSIS + CTCF gene family and encodes a transcriptional regulator protein with 11 highly conserved zinc finger (ZF) domains. Mutations in this gene have been associated with invasive breast cancers, prostate cancers, and Wilms’ tumors.
fancd2	FA Complementation Group D2: Part of the Fanconi anemia complementation group (FANC) and required to maintain chromosomal stability. Plays a role in preventing breakage and loss of mis segregating chromatin at the end of cell division. Associated with Fanconi Anemia.
hsd17b11	Hydroxysteroid 17-Beta Dehydrogenase 11: A short-chain alcohol dehydrogenases which metabolizes secondary alcohols and ketones. Associated with Cutaneous T Cell Lymphoma and Lymphoma.

Appendix B: Clinical Variables – Factor Analysis

Appendix B1:
PFA with 4 Factors

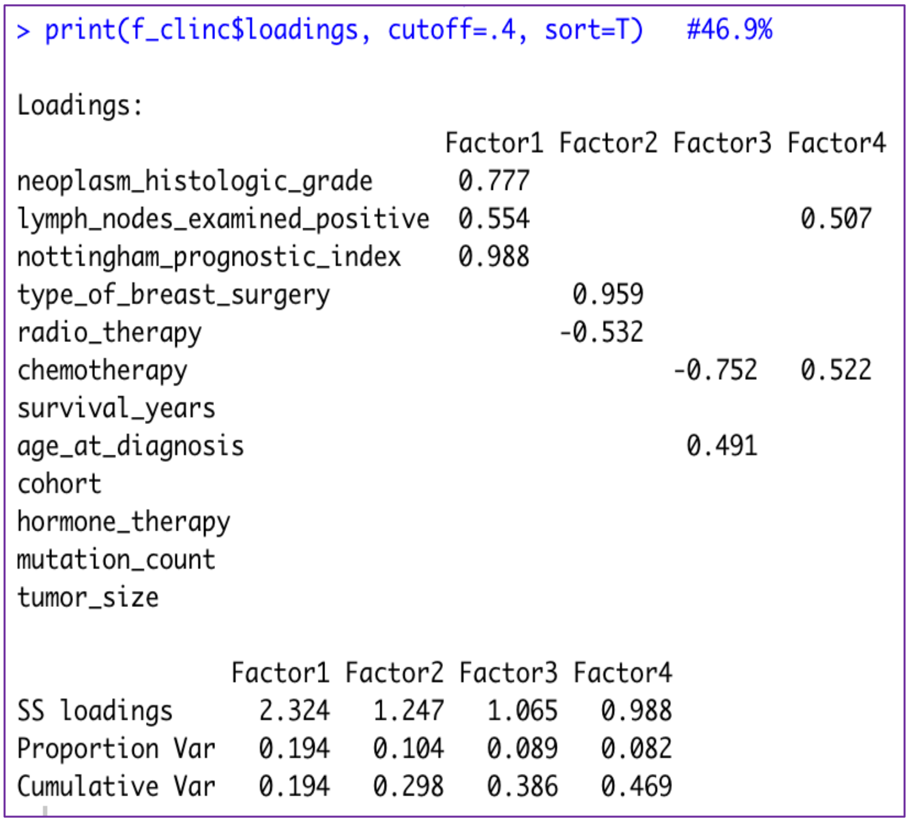


```
> print(pr_clinc$loadings, cutoff=.4, sort=T)   #61.6%

Loadings:
                              RC1    RC2    RC3    RC4   
neoplasm_histologic_grade      0.672                     
lymph_nodes_examined_positive  0.698                     
nottingham_prognostic_index    0.925                     
tumor_size                     0.557                     
age_at_diagnosis                      0.838              
chemotherapy                   0.515 -0.619              
hormone_therapy                       0.638              
cohort                                       0.787       
mutation_count                               0.739       
type_of_breast_surgery                              -0.832
radio_therapy                                        0.829
survival_years                -0.426                     

                 RC1   RC2   RC3   RC4
SS loadings    2.747 1.605 1.539 1.498
Proportion Var 0.229 0.134 0.128 0.125
Cumulative Var 0.229 0.363 0.491 0.616
```



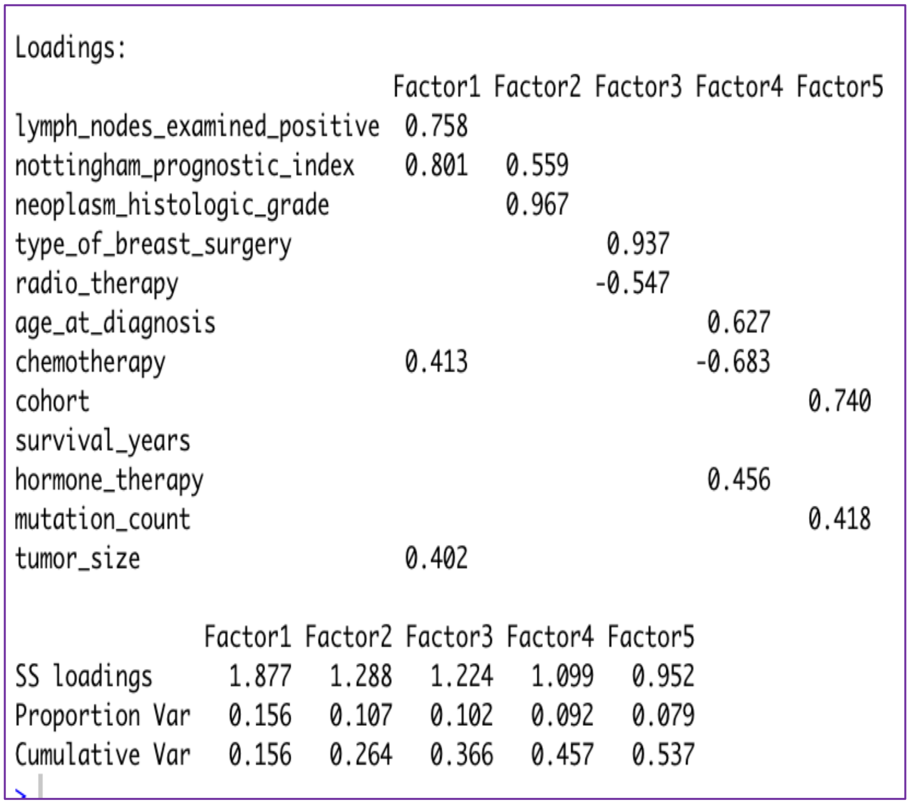
Appendix B2:
Horn's Parallel Analysis

Appendix B3: CFA with 4 Factors



```
> print(f_clinc$loadings, cutoff=.4, sort=T)   #46.9%

Loadings:
                              Factor1 Factor2 Factor3 Factor4
neoplasm_histologic_grade      0.777                         
lymph_nodes_examined_positive  0.554                   0.507 
nottingham_prognostic_index    0.988                         
type_of_breast_surgery                 0.959                 
radio_therapy                         -0.532                 
chemotherapy                                  -0.752   0.522 
survival_years                                               
age_at_diagnosis                               0.491         
cohort                                                       
hormone_therapy                                              
mutation_count                                               
tumor_size                                                   

               Factor1 Factor2 Factor3 Factor4
SS loadings      2.324   1.247   1.065   0.988
Proportion Var   0.194   0.104   0.089   0.082
Cumulative Var   0.194   0.298   0.386   0.469
```

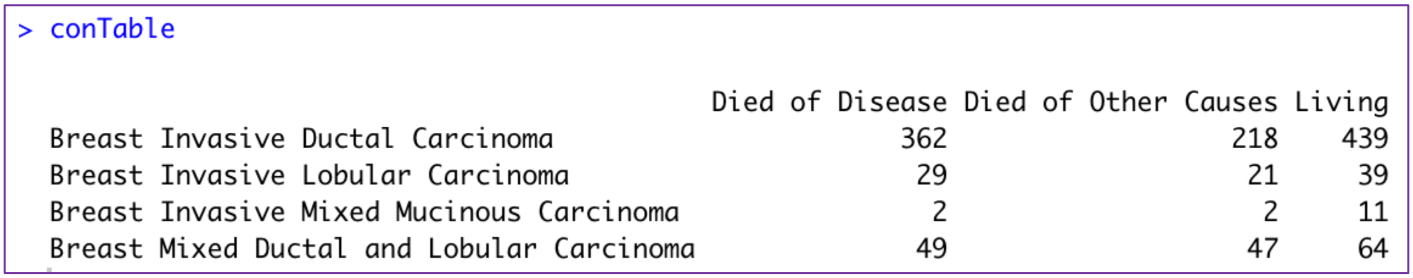
Appendix B4: CFA with 5 Factors



```
Loadings:
                              Factor1 Factor2 Factor3 Factor4 Factor5
lymph_nodes_examined_positive  0.758                                 
nottingham_prognostic_index    0.801   0.559                         
neoplasm_histologic_grade              0.967                         
type_of_breast_surgery                         0.937                 
radio_therapy                                 -0.547                 
age_at_diagnosis                                       0.627         
chemotherapy                   0.413                  -0.683         
cohort                                                         0.740 
survival_years                                                       
hormone_therapy                                        0.456         
mutation_count                                                 0.418 
tumor_size                     0.402                                 

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      1.877   1.288   1.224   1.099   0.952
Proportion Var   0.156   0.107   0.102   0.092   0.079
Cumulative Var   0.156   0.264   0.366   0.457   0.537
>
```

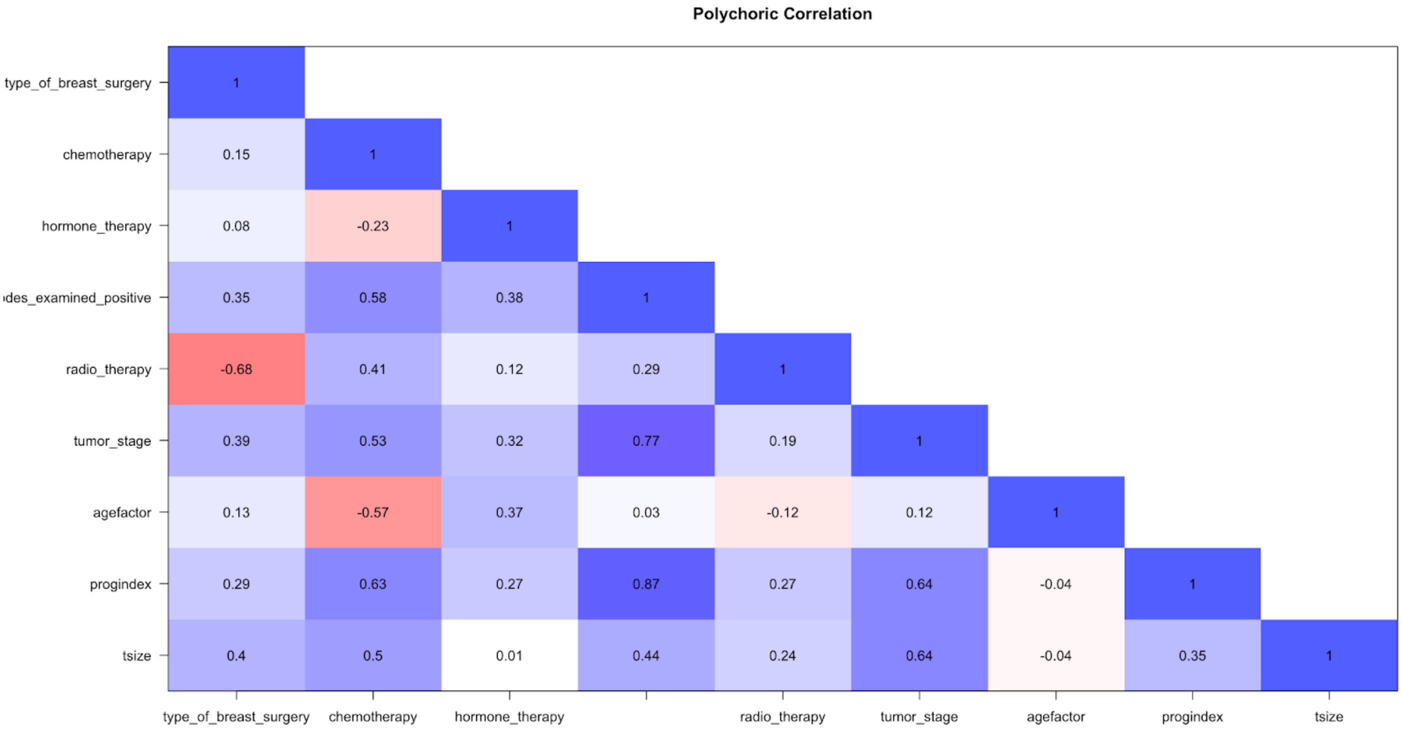
Appendix C: Clinical Variables – Correspondence Analysis and Polychoric Correlation Analysis

Appendix C1: Contingency Table for Categorical Variables



```
> conTable

                                           Died of Disease Died of Other Causes Living
  Breast Invasive Ductal Carcinoma                     362                  218    439
  Breast Invasive Lobular Carcinoma                     29                   21     39
  Breast Invasive Mixed Mucinous Carcinoma               2                    2     11
  Breast Mixed Ductal and Lobular Carcinoma             49                   47     64
```

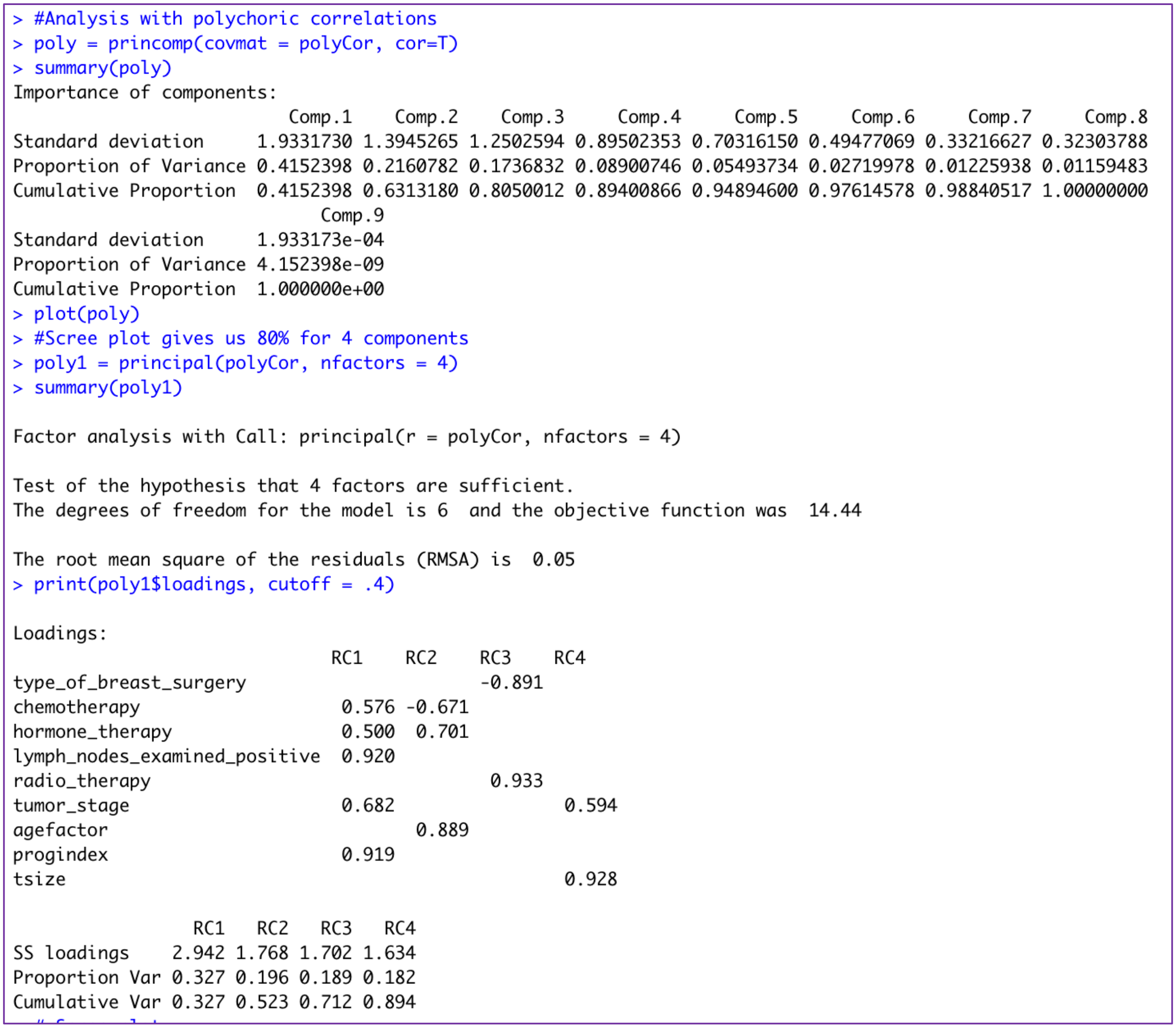
Appendix C2: Polychoric Correlations



Polychoric Correlation

	type_of_breast_surgery	chemotherapy	hormone_therapy		radio_therapy	tumor_stage	agefactor	progindex	tsize
type_of_breast_surgery	1								
chemotherapy	0.15	1							
hormone_therapy	0.08	-0.23	1						
ıdes_examined_positive	0.35	0.58	0.38	1					
radio_therapy	-0.68	0.41	0.12	0.29	1				
tumor_stage	0.39	0.53	0.32	0.77	0.19	1			
agefactor	0.13	-0.57	0.37	0.03	-0.12	0.12	1		
progindex	0.29	0.63	0.27	0.87	0.27	0.64	-0.04	1	
tsize	0.4	0.5	0.01	0.44	0.24	0.64	-0.04	0.35	1

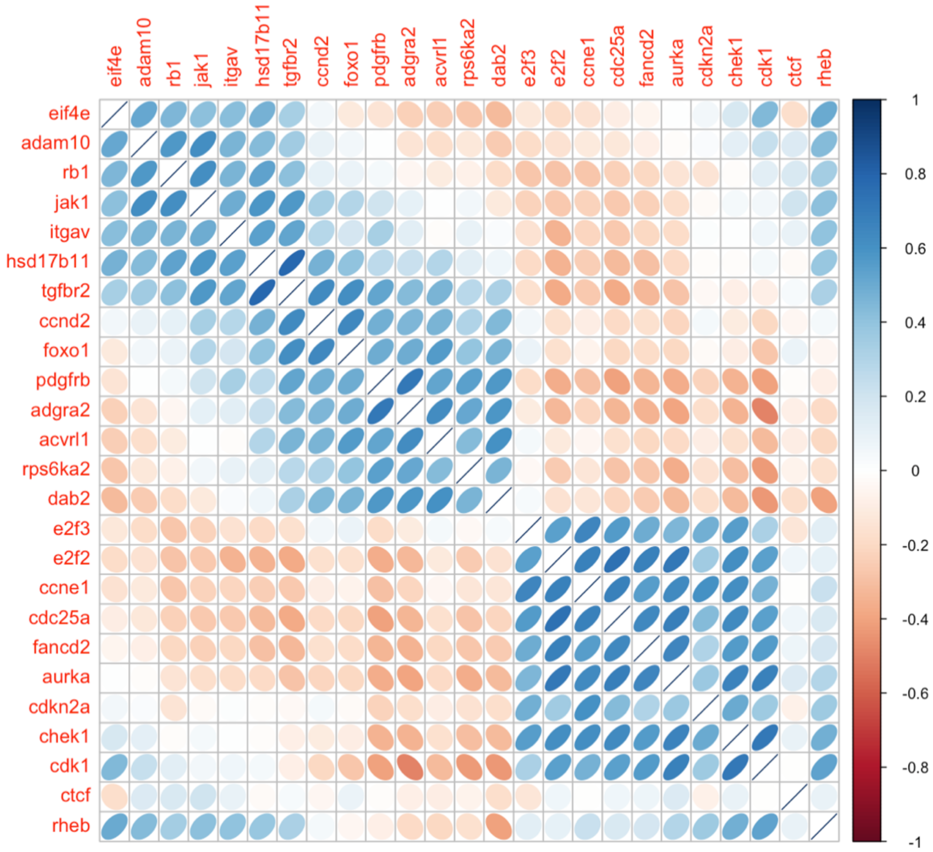
Appendix C3: Polychoric Factor Analysis



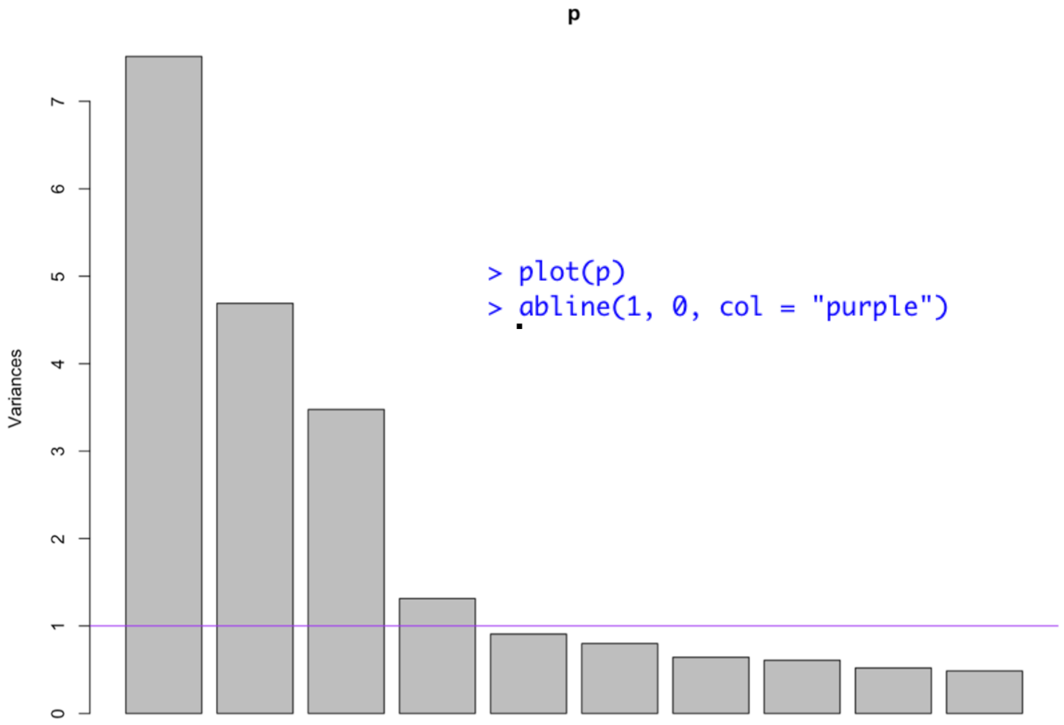
```
> #Analysis with polychoric correlations
> poly = princomp(covmat = polyCor, cor=T)
> summary(poly)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     1.9331730 1.3945265 1.2502594 0.89502353 0.70316150 0.49477069 0.33216627 0.32303788
Proportion of Variance 0.4152398 0.2160782 0.1736832 0.08900746 0.05493734 0.02719978 0.01225938 0.01159483
Cumulative Proportion  0.4152398 0.6313180 0.8050012 0.89400866 0.94894600 0.97614578 0.98840517 1.00000000
                             Comp.9
Standard deviation     1.933173e-04
Proportion of Variance 4.152398e-09
Cumulative Proportion  1.000000e+00
> plot(poly)
> #Scree plot gives us 80% for 4 components
> poly1 = principal(polyCor, nfactors = 4)
> summary(poly1)

Factor analysis with Call: principal(r = polyCor, nfactors = 4)

Test of the hypothesis that 4 factors are sufficient.
The degrees of freedom for the model is 6  and the objective function was  14.44

The root mean square of the residuals (RMSA) is  0.05
> print(poly1$loadings, cutoff = .4)

Loadings:
                              RC1    RC2    RC3    RC4   
type_of_breast_surgery                      -0.891       
chemotherapy                   0.576 -0.671              
hormone_therapy                0.500  0.701              
lymph_nodes_examined_positive  0.920                     
radio_therapy                                0.933       
tumor_stage                    0.682                0.594
agefactor                             0.889              
progindex                      0.919                     
tsize                                               0.928

                 RC1   RC2   RC3   RC4
SS loadings    2.942 1.768 1.702 1.634
Proportion Var 0.327 0.196 0.189 0.182
Cumulative Var 0.327 0.523 0.712 0.894
```

Appendix D: Genomic Variables – Factor Analysis

Appendix D1: Correlation Plot with all 25 genomic variables – Angle of Eigenvector Method

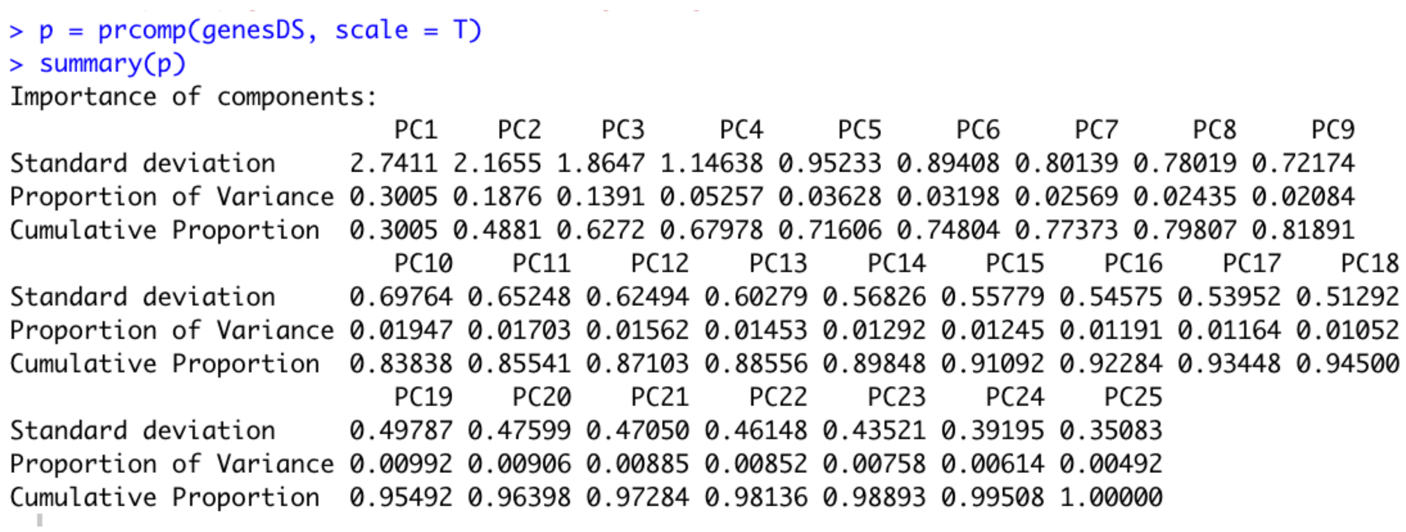


Appendix D2: Scree Plot of Principal Components



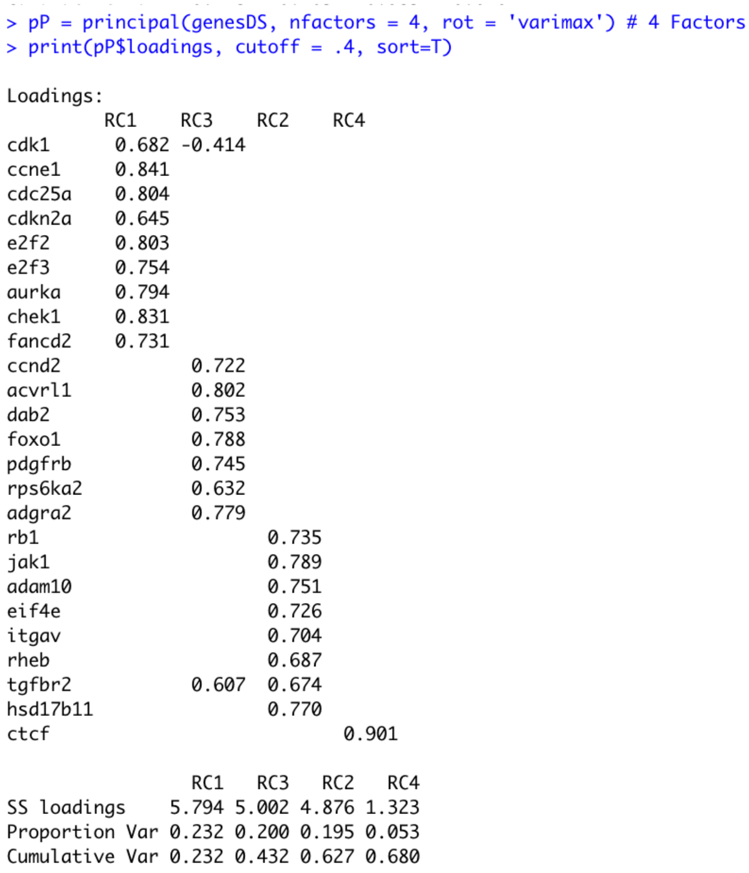
```
> plot(p)
> abline(1, 0, col = "purple")
```

Appendix D3: Summary of Principal Components



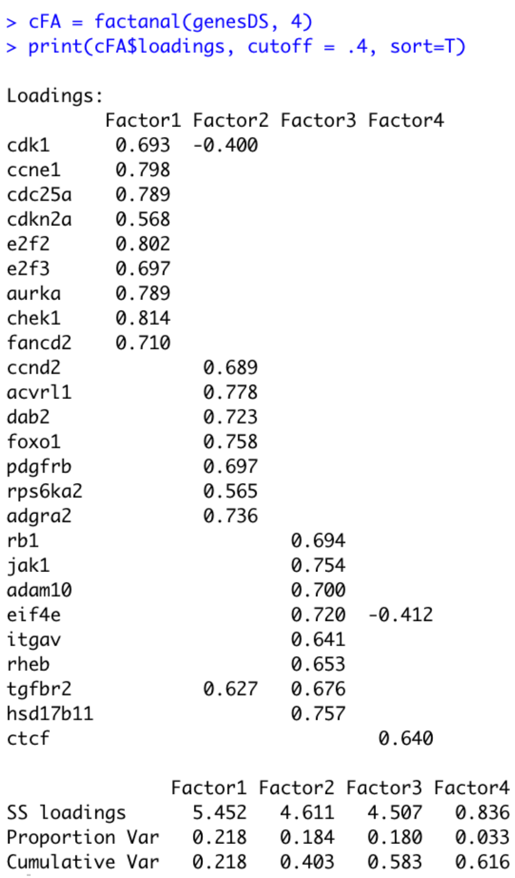
```
> p = prcomp(genesDS, scale = T)
> summary(p)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.7411 2.1655 1.8647 1.14638 0.95233 0.89408 0.80139 0.78019 0.72174
Proportion of Variance 0.3005 0.1876 0.1391 0.05257 0.03628 0.03198 0.02569 0.02435 0.02084
Cumulative Proportion  0.3005 0.4881 0.6272 0.67978 0.71606 0.74804 0.77373 0.79807 0.81891
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
Standard deviation     0.69764 0.65248 0.62494 0.60279 0.56826 0.55779 0.54575 0.53952 0.51292
Proportion of Variance 0.01947 0.01703 0.01562 0.01453 0.01292 0.01245 0.01191 0.01164 0.01052
Cumulative Proportion  0.83838 0.85541 0.87103 0.88556 0.89848 0.91092 0.92284 0.93448 0.94500
                          PC19    PC20    PC21    PC22    PC23    PC24    PC25
Standard deviation     0.49787 0.47599 0.47050 0.46148 0.43521 0.39195 0.35083
Proportion of Variance 0.00992 0.00906 0.00885 0.00852 0.00758 0.00614 0.00492
Cumulative Proportion  0.95492 0.96398 0.97284 0.98136 0.98893 0.99508 1.00000
```

Appendix D4: PFA with 4 Factors



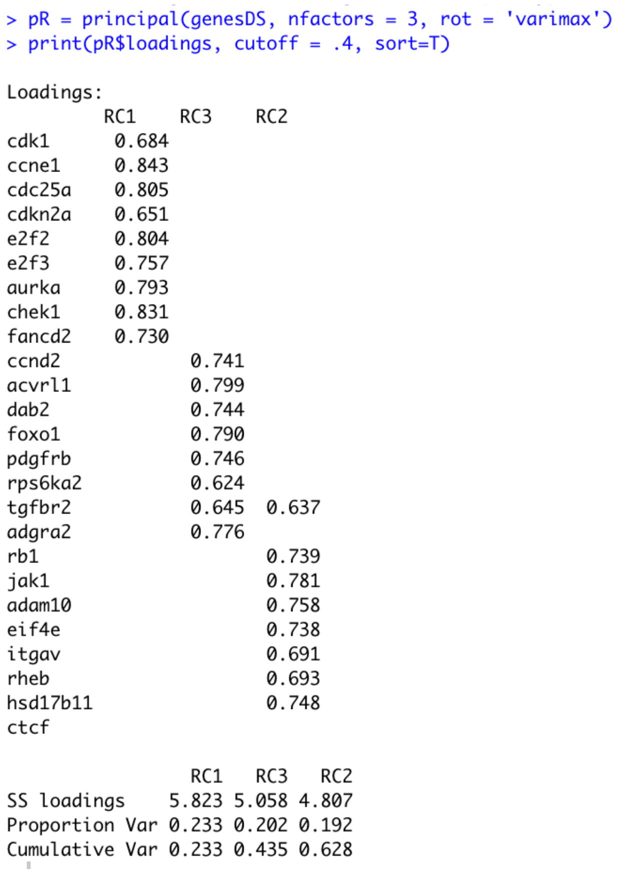
```
> pP = principal(genesDS, nfactors = 4, rot = 'varimax') # 4 Factors
> print(pP$loadings, cutoff = .4, sort=T)

Loadings:
         RC1    RC3    RC2    RC4   
cdk1      0.682 -0.414              
ccne1     0.841                     
cdc25a    0.804                     
cdkn2a    0.645                     
e2f2      0.803                     
e2f3      0.754                     
aurka     0.794                     
chek1     0.831                     
fancd2    0.731                     
ccnd2            0.722              
acvrl1           0.802              
dab2             0.753              
foxo1            0.788              
pdgfrb           0.745              
rps6ka2          0.632              
adgra2           0.779              
rb1                     0.735       
jak1                    0.789       
adam10                  0.751       
eif4e                   0.726       
itgav                   0.704       
rheb                    0.687       
tgfbr2           0.607  0.674       
hsd17b11                0.770       
ctcf                           0.901

                 RC1   RC3   RC2   RC4
SS loadings    5.794 5.002 4.876 1.323
Proportion Var 0.232 0.200 0.195 0.053
Cumulative Var 0.232 0.432 0.627 0.680
```

Appendix D5: CFA with 4 Factors

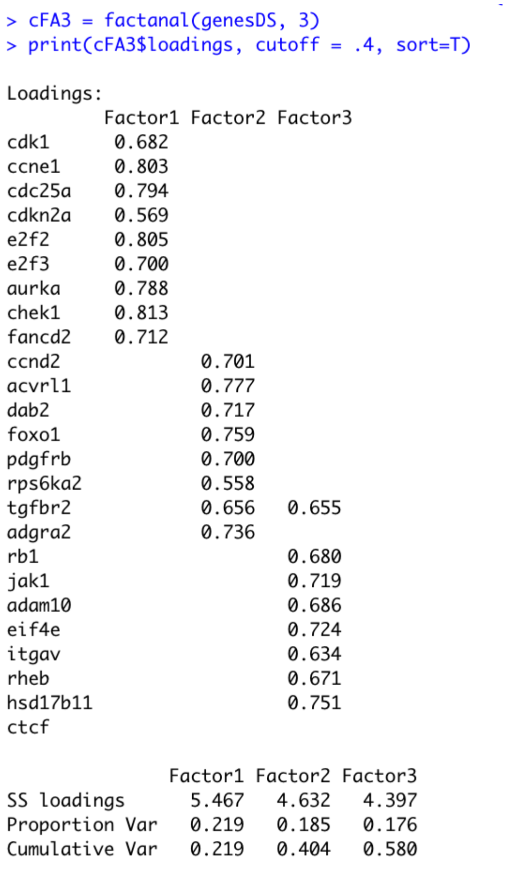


```
> cFA = factanal(genesDS, 4)
> print(cFA$loadings, cutoff = .4, sort=T)

Loadings:
         Factor1 Factor2 Factor3 Factor4
cdk1      0.693  -0.400                 
ccne1     0.798                         
cdc25a    0.789                         
cdkn2a    0.568                         
e2f2      0.802                         
e2f3      0.697                         
aurka     0.789                         
chek1     0.814                         
fancd2    0.710                         
ccnd2             0.689                 
acvrl1            0.778                 
dab2              0.723                 
foxo1             0.758                 
pdgfrb            0.697                 
rps6ka2           0.565                 
adgra2            0.736                 
rb1                       0.694         
jak1                      0.754         
adam10                    0.700         
eif4e                     0.720  -0.412 
itgav                     0.641         
rheb                      0.653         
tgfbr2            0.627   0.676         
hsd17b11                  0.757         
ctcf                              0.640 

               Factor1 Factor2 Factor3 Factor4
SS loadings      5.452   4.611   4.507   0.836
Proportion Var   0.218   0.184   0.180   0.033
Cumulative Var   0.218   0.403   0.583   0.616
```

Appendix D6: PFA and CFA with 3 Factors Comparison

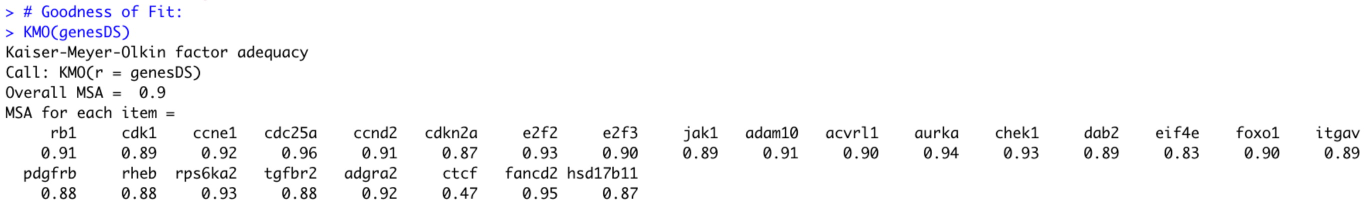


```
> pR = principal(genesDS, nfactors = 3, rot = 'varimax')
> print(pR$loadings, cutoff = .4, sort=T)

Loadings:
         RC1    RC3    RC2   
cdk1     0.684               
ccne1    0.843               
cdc25a   0.805               
cdkn2a   0.651               
e2f2     0.804               
e2f3     0.757               
aurka    0.793               
chek1    0.831               
fancd2   0.730               
ccnd2           0.741        
acvrl1          0.799        
dab2            0.744        
foxo1           0.790        
pdgfrb          0.746        
rps6ka2         0.624        
tgfbr2          0.645  0.637 
adgra2          0.776        
rb1                    0.739 
jak1                   0.781 
adam10                 0.758 
eif4e                  0.738 
itgav                  0.691 
rheb                   0.693 
hsd17b11               0.748 
ctcf                         

                 RC1   RC3   RC2
SS loadings    5.823 5.058 4.807
Proportion Var 0.233 0.202 0.192
Cumulative Var 0.233 0.435 0.628
```



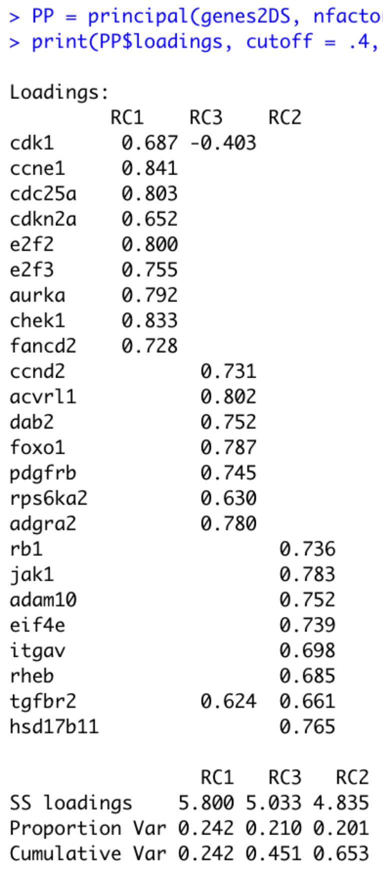
```
> cFA3 = factanal(genesDS, 3)
> print(cFA3$loadings, cutoff = .4, sort=T)

Loadings:
         Factor1 Factor2 Factor3
cdk1     0.682                  
ccne1    0.803                  
cdc25a   0.794                  
cdkn2a   0.569                  
e2f2     0.805                  
e2f3     0.700                  
aurka    0.788                  
chek1    0.813                  
fancd2   0.712                  
ccnd2            0.701          
acvrl1           0.777          
dab2             0.717          
foxo1            0.759          
pdgfrb           0.700          
rps6ka2          0.558          
tgfbr2           0.656   0.655  
adgra2           0.736          
rb1                      0.680  
jak1                     0.719  
adam10                   0.686  
eif4e                    0.724  
itgav                    0.634  
rheb                     0.671  
hsd17b11                 0.751  
ctcf                            

               Factor1 Factor2 Factor3
SS loadings      5.467   4.632   4.397
Proportion Var   0.219   0.185   0.176
Cumulative Var   0.219   0.404   0.580
```

Appendix D7: KMO Goodness of Fit

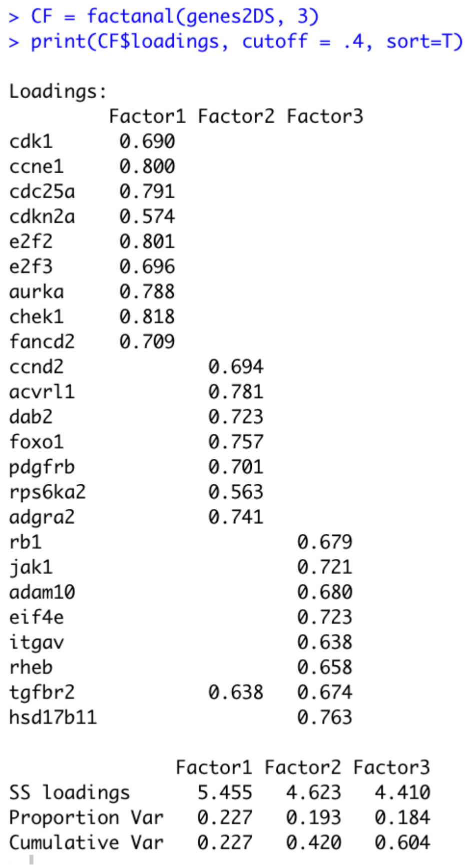


```
> # Goodness of Fit:
> KMO(genesDS)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = genesDS)
Overall MSA =  0.9
MSA for each item = 
     rb1     cdk1    ccne1   cdc25a    ccnd2   cdkn2a     e2f2     e2f3     jak1   adam10   acvrl1    aurka    chek1     dab2    eif4e    foxo1    itgav 
    0.91     0.89     0.92     0.96     0.91     0.87     0.93     0.90     0.89     0.91     0.90     0.94     0.93     0.89     0.83     0.90     0.89 
  pdgfrb     rheb  rps6ka2   tgfbr2   adgra2     ctcf   fancd2 hsd17b11 
    0.88     0.88     0.93     0.88     0.92     0.47     0.95     0.87 
```

Appendix D8: PFA and CFA without ctcf, with 4 Factors



```
> PP = principal(genes2DS, nfactors = 3, rot = 'varimax')
> print(PP$loadings, cutoff = .4, sort=T)

Loadings:
         RC1    RC3    RC2   
cdk1      0.687 -0.403       
ccne1     0.841              
cdc25a    0.803              
cdkn2a    0.652              
e2f2      0.800              
e2f3      0.755              
aurka     0.792              
chek1     0.833              
fancd2    0.728              
ccnd2            0.731       
acvrl1           0.802       
dab2             0.752       
foxo1            0.787       
pdgfrb           0.745       
rps6ka2          0.630       
adgra2           0.780       
rb1                     0.736
jak1                    0.783
adam10                  0.752
eif4e                   0.739
itgav                   0.698
rheb                    0.685
tgfbr2           0.624  0.661
hsd17b11                0.765

                 RC1   RC3   RC2
SS loadings    5.800 5.033 4.835
Proportion Var 0.242 0.210 0.201
Cumulative Var 0.242 0.451 0.653
```



```
> CF = factanal(genes2DS, 3)
> print(CF$loadings, cutoff = .4, sort=T)

Loadings:
         Factor1 Factor2 Factor3
cdk1     0.690                  
ccne1    0.800                  
cdc25a   0.791                  
cdkn2a   0.574                  
e2f2     0.801                  
e2f3     0.696                  
aurka    0.788                  
chek1    0.818                  
fancd2   0.709                  
ccnd2            0.694          
acvrl1           0.781          
dab2             0.723          
foxo1            0.757          
pdgfrb           0.701          
rps6ka2          0.563          
adgra2           0.741          
rb1                      0.679  
jak1                     0.721  
adam10                   0.680  
eif4e                    0.723  
itgav                    0.638  
rheb                     0.658  
tgfbr2           0.638   0.674  
hsd17b11                 0.763  

               Factor1 Factor2 Factor3
SS loadings      5.455   4.623   4.410
Proportion Var   0.227   0.193   0.184
Cumulative Var   0.227   0.420   0.604
```

Appendix D9: RC1 – Genes that are CDK/cyclin and proteins in the kinase family that when unregulated associated with the uncontrolled proliferation of cancer cells. *All descriptions from GeneCards*.

cdk1: Cyclin Dependent Kinase 1: Protein encoded by this gene is part of the Ser/Thr protein kinase family. Catalytic subunit of highly conserved kinase complex, essential for G1/S and G2/M phase transition of eukaryotic cell cycle. Associated with Retinoblastoma and Breast Cancer.

ccne1: Cyclin E1: Protein encoded by this gene is part of the highly conserved cyclin family. Cyclins function as regulators of CDK kinases. Overexpression of this gene has been observed in many tumors. Associated with clear cell Adenocarcinoma of the Ovary and Retinoblastoma.

cdc25a: Cell Division Cycle 25A: Part of the CDC25 family of phosphatases. Required for progression from G1 to the S phase of the cell cycle. Activates the cyclin-dependent kinase CDC2. Associated with Retinoblastoma and Ataxia-telangiectasia.

cdkn2a: Cyclin Dependent Kinase Inhibitor 2A: Gene generates several transcript variants which different in their first exons. Gene is frequently mutated or deleted in a wide variety of tumors and is known to be an important tumor suppressor gene. Loss of this gene is shown to be significant in many cancers.

e2f2: E2F Transcription Factor 2: Protein encoded by this gene is a part of the E2F family of transcription factors. Plays a crucial role in control of cell cycle and action of tumor suppressor proteins. Associated with Retinoblastoma and Encapsulated Thymoma.

e2f3: E2F Transcription Factor 3: Encodes a member of a small family of transcription factors that function through binding of DP interaction partner proteins. Associated with Retinoblastoma and Bladder Cancer.

aurka: Aurora Kinase A: Protein encoded by this gene is a cell cycle-regulated kinase that appears to be involved in microtubule formation and/or stabilization at the spindle pole during chromosome segregation. Gene may play a role in tumor development and progression. Associated with Colorectal Cancer.

chek1: Checkpoint Kinase 1: Protein encoded by this gene is part of the Ser/Th protein kinase family. Required for checkpoint cell cycle arrest in response to DNA damage or the presence of replicated DNA. Associated with Ataxia-Telangiectasia and Li-Fraumeni Syndrome.

fancd2: FA Complementation Group D2: Part of the Fanconi anemia complementation group (FANC) and required to maintain chromosomal stability. Plays a role in preventing breakage and loss of mis segregating chromatin at the end of cell division. Associated with Fanconi Anemia.

Appendix D10: RC3 – Genes that are receptor, encoder, or adaptor proteins that serve the serine-threonine kinase domain or subdomains. *All descriptions from GeneCards*

ccnd2: Cyclin D2: Protein coded by this gene belongs to the highly conserved cyclin family. Cyclins function as regulators of CDK kinases. Has been shown in many cancer types if unregulated.

acvrl1: Activin A Receptor Like Type 1: Encodes a type 1 cell-surface receptor for the TGF-beta superfamily of ligands. Shares high degree of similarity. To the serine-threonine kinase subdomains. Associated with Telangiectasia and Hereditary Hemorrhagic.

dab2: DAB Adaptor Protein 2: Encodes a mitogen-responsive phosphoprotein. Expressed in a normal ovarian epithelial cell but is downregulated or absent from ovarian carcinoma cell lines, suggesting its role as a tumor suppressor. Associated with Teratocarcinoma.

foxo1: Forkhead Box O1: Part of the forkhead family of transcription factors. Specific function has not yet been determined, but it may play a role in myogenic growth and differentiation. Associated with Rhabdomyosarcoma and Glioma.

pdgfrb: Platelet Derived Growth Factor Receptor Beta: Protein encoded by this gene is a cell surface tyrosine kinase receptor for the platelet-derived growth factor family. This gene is essential for the normal development of the cardiovascular system and aids in rearrangement of the actin cytoskeleton. Associated with Premature Aging Syndrome and Kosaki Overgrowth Syndrome.

rps6ka2: Ribosomal Protein S6 Kinase A2: Encodes a member of the RSK (ribosomal S6 kinase) family of the serine/threonine kinases. Activity of this protein has been implicated in controlling cell growth and differentiation. Associated with Coffin-Lowry Syndrome and Autism.

tgfbr2: Transforming Growth Factor Beta Receptor 2: The protein encoded by this gene is a transmembrane protein that has a protein kinase domain and forms a heterodimeric complex with TGF-beta receptor type-1 and binds TGF-beta. Mutations of this gene have been associated with Marfan Syndrome, Loeys-Deitz Aortic Aneurism Syndrome, and the develop of various types of tumors. Diseases associated with TGFBR2 include Loeys-Dietz Syndrome, Colorectal Cancer, and Hereditary Nonpolyposis.

adgra2: Adhesion G Protein-Coupled Receptor A2: Part of the adhesion-GPCR family of receptors. Endothelial receptor which functions together with RECK

Appendix D11: RC2 – Genes that are either tumor suppressors, binding proteins, or proteins that can stabilize or regulate cancer progression. *All descriptions from GeneCards.*

rb1: RB Transcriptional Corepressor 1: Protein encoded by this gene is a negative regulator of the cell cycle and was the first tumor suppressor gene found. Associated with Retinoblastoma and Small Cell Lung Cancer.

jak1: Janus Kinase 1: Encodes a membrane protein that is part of a class of protein-tyrosine kinases (PTK). Gene plays a crucial role in effecting the expression of genes that mediate inflammation, epithelial remodeling, and metastatic cancer progression. Associated with Autoinflammation, Immune Dysregulation, and Eosinophilia.

adam10: ADAM Metallopeptidase Domain 10: Cell surface proteins with a unique structure possessing both potential adhesion and protease domains. Associated with Reticulate Pigmentation of Kitamura and Alzheimer Disease.

itgav: Integrin Subunit Alpha V: Product of this gene belongs to the integrin alpha chain family. Integrins are heterodimeric integral membrane proteins and may regulate angiogenesis and cancer progression. Associated with West Nile Virus and Herpes Simplex.

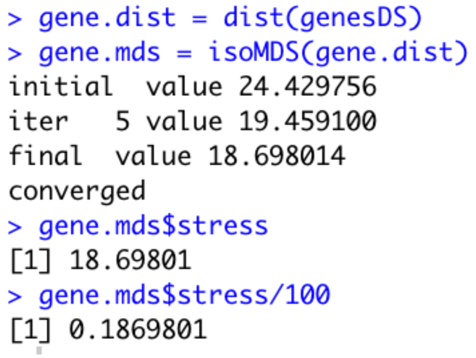
rheb: RAS Homolog, HTORC1 Binding: Gene is a member of a small GTPase superfamily and encodes a lipid-anchored, cell membrane protein with five repeats of the RAS-related GTP-binding region. Associated with Tuberous Sclerosis and Hemimegaloencephaly.

tgfbr2: Transforming Growth Factor Beta Receptor 2: The protein encoded by this gene is a transmembrane protein that has a protein kinase domain and forms a heterodimeric complex with TGF-beta receptor type-1 and binds TGF-beta. Mutations of this gene have been associated with Marfan Syndrome, Loeys-Deitz Aortic Aneurism Syndrome, and the develop of various types of tumors. Diseases associated with TGFBR2 include Loeys-Dietz Syndrome, Colorectal Cancer, and Hereditary Nonpolyposis.

hsd17b11: Hydroxysteroid 17-Beta Dehydrogenase 11: A short-chain alcohol dehydrogenases which metabolizes secondary alcohols and ketones. Associated with Cutaneous T Cell Lymphoma and Lymphoma

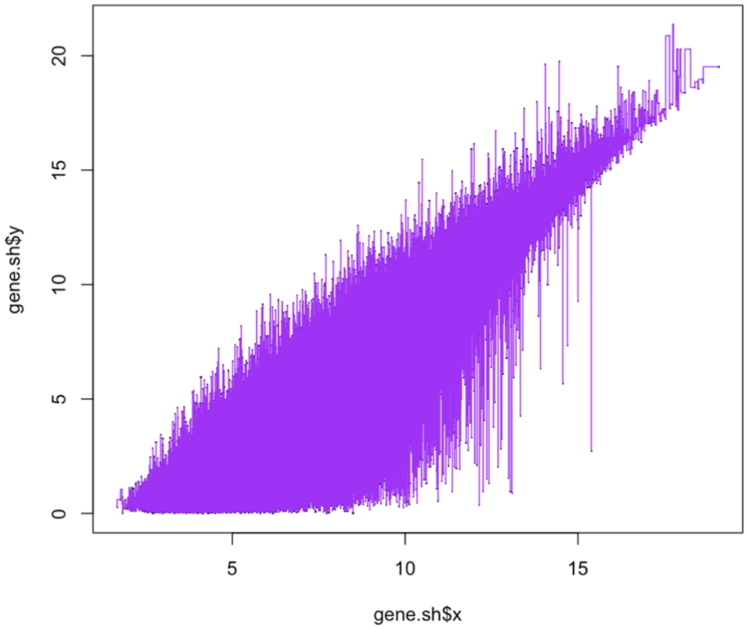
Appendix E: Genomic Variables – MDS and Cluster Analysis

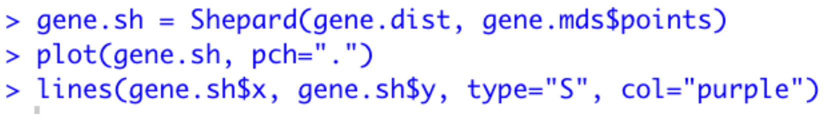
Appendix E1: Multidimensional Scaling (MDS)



```
> gene.dist = dist(genesDS)
> gene.mds = isoMDS(gene.dist)
initial  value 24.429756 
iter   5 value 19.459100
final  value 18.698014 
converged
> gene.mds$stress
[1] 18.69801
> gene.mds$stress/100
[1] 0.1869801
```

Appendix E2: Shepard's Diagram and Stress (Kruskal's) Function

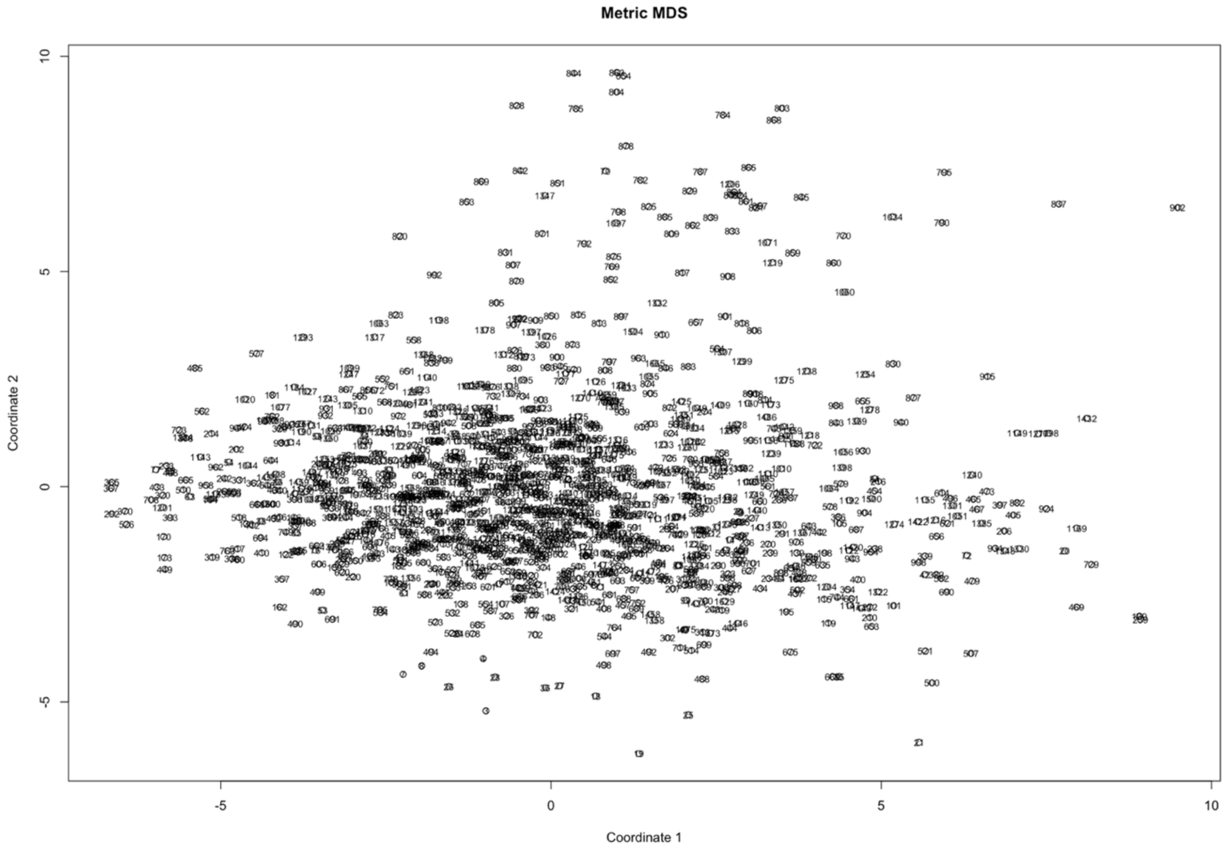




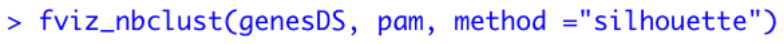
```
> gene.sh = Shepard(gene.dist, gene.mds$points)
> plot(gene.sh, pch=".")
> lines(gene.sh$x, gene.sh$y, type="S", col="purple")
```

Appendix E3: Multidimensional Scaling (MDS) Plot

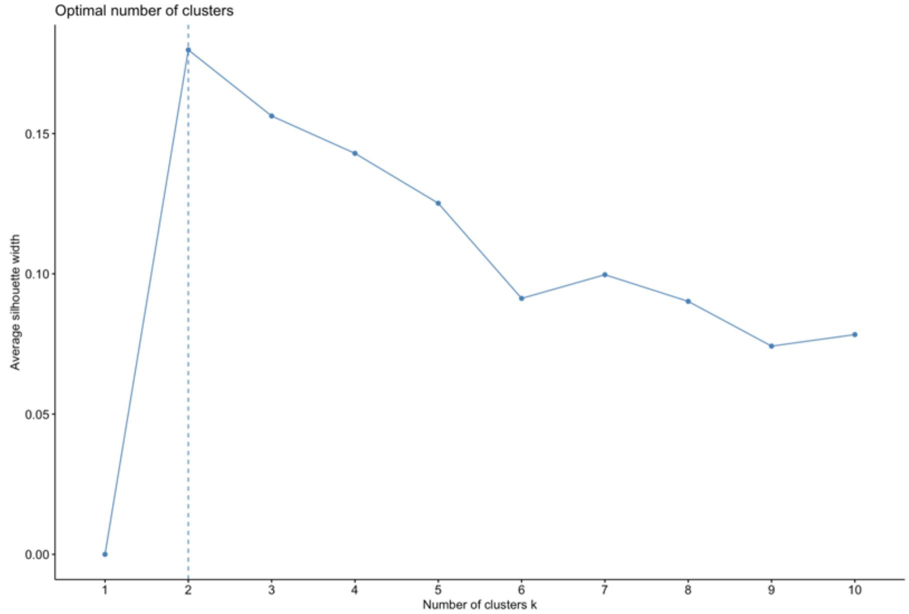
```
> d = dist(genesDS)
> fit = cmdscale(d, eig=TRUE, k=2)
> #plot fit
> x = fit$points[,1]
> y = fit$points[,2]
> plot(x, y, xlab="Coordinate 1", ylab="Coordinate 2", main="Metric MDS")
> text(x, y, labels = row.names(genesDS), cex=.7)
```



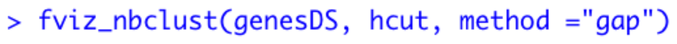
Appendix E4: Average Silhouette Method



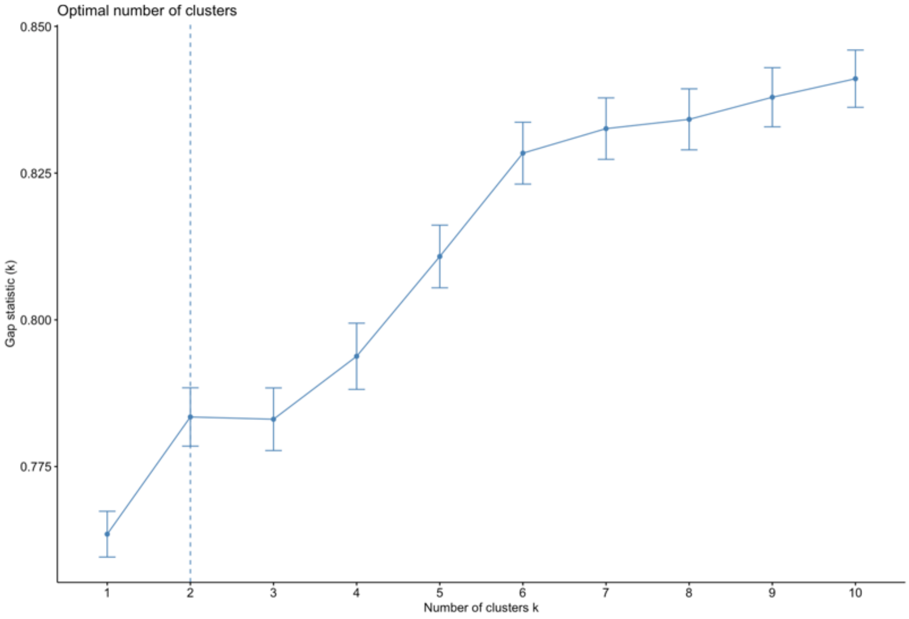
```
> fviz_nbclust(genesDS, pam, method ="silhouette")
```



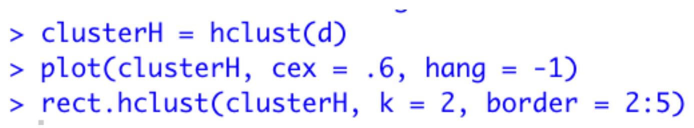
Appendix E5: Gap Statistics Method



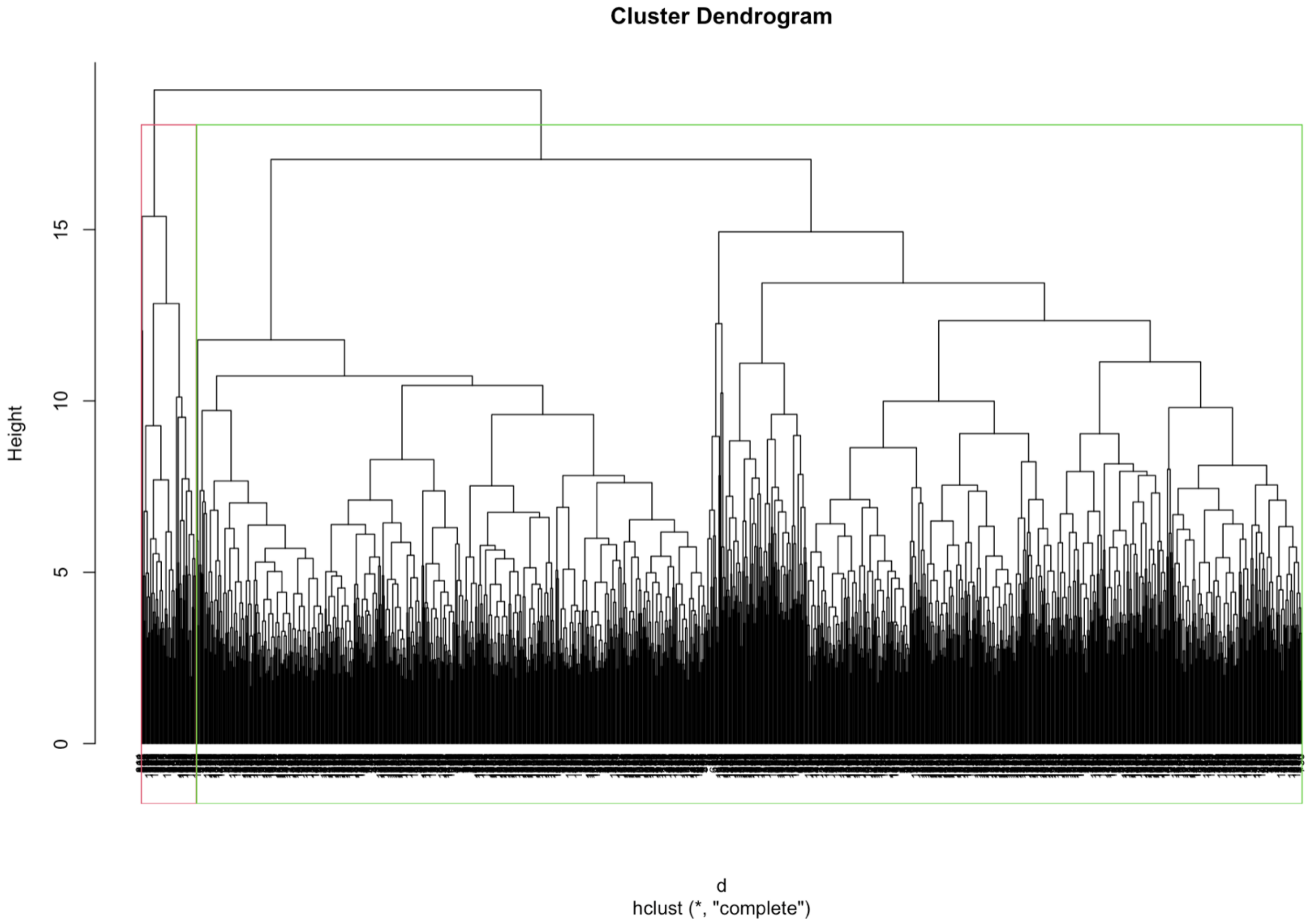
```
> fviz_nbclust(genesDS, hcut, method ="gap")
```



Appendix E6: Dendrogram for Hierarchical Clustering

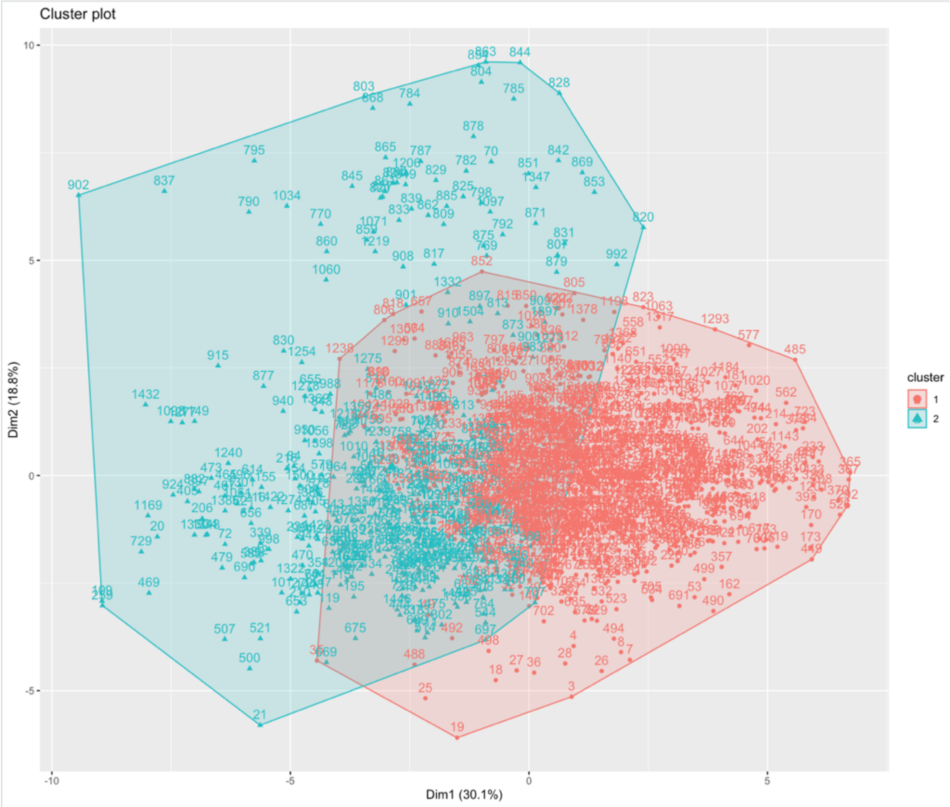


```
> clusterH = hclust(d)
> plot(clusterH, cex = .6, hang = -1)
> rect.hclust(clusterH, k = 2, border = 2:5)
```

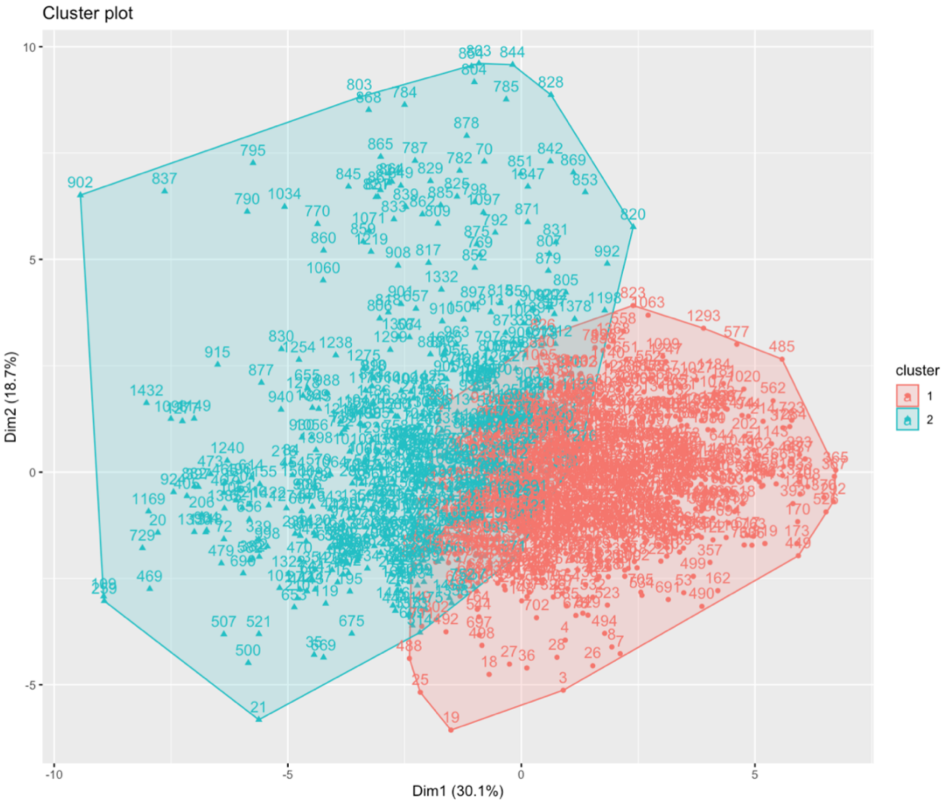


```
> geneCut = hcut(genesDS, k = 2, stand=TRUE)
> fviz_cluster(geneCut, data = genesDS)
```

Appendix E7: H-Cluster Plot, k =2



```
> genePam = pam(genesDS, k = 2)
> fviz_cluster(genePam, data = genesDS)
```



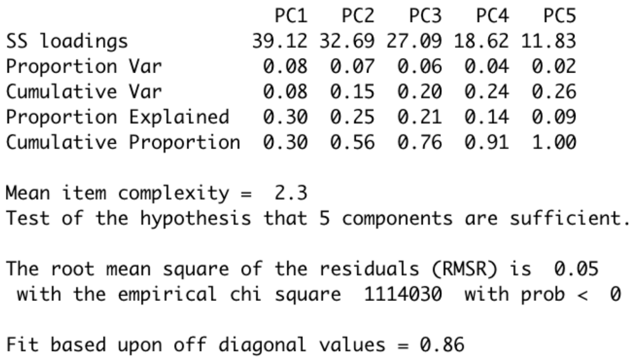
Appendix E8: K-medoids (PAM-Cluster) Plot, k = 2

Appendix G: Cody Le

Cody’s role in the project included planning and organizing the zoom meetings, discussion, and progress of the milestones which the team used google drive and workspace to collaborate virtually. The team met weekly to discuss each milestone, which was divided into individual tasks, which was reviewed throughout the week and finalized at each subsequent meeting. In the exploration of the dataset, Cody focused on cleaning and preparing the data in R, specifically performing the dimensionality reduction on the genomic variables using principal component analysis. For the in-depth analysis, Cody further focused on the genomic variables by researching the genes, gene functions, and documented mutations and diseases associated with the genes. Cody performed principal factor analysis, common factor analysis, and cluster analysis on the genomic variables. Lastly, Cody organized the google slides for the video presentation, formatted the sides, and prepared the introduction and data preparation sections of the presentation.

The original dataset had 693 columns of which 30 were clinical variables and 663 genomic variables. Cleaning was performed on the genomic variables to remove columns with all zero values or missing values. The remaining 488 genomic variables were reduced through principal component analysis and further reduced through principal factor analysis. The genomic variables were already normalized through z-score normalization and the dimensionality reduction proved to work well with the data as it optimized the dimensions and remove the majority of the columns while keeping acceptable cumulative variance. Due to the high number of dimensions, principal factor analysis to rotate the loadings and select the significant variables was a little challenging at the beginning. It was decided that five factors would be chosen based on the knee from the initial principal component analysis. The loadings showed that at four components, 91% of the variance was captured. Sorting through each of the components, the team set a threshold for selecting the variables based on the loadings, in this case, loadings with 0.6 or higher was selected. The result was 25 variables from the first four components.

Principal Factor Analysis (PFA) with Varimax, 5 Factors



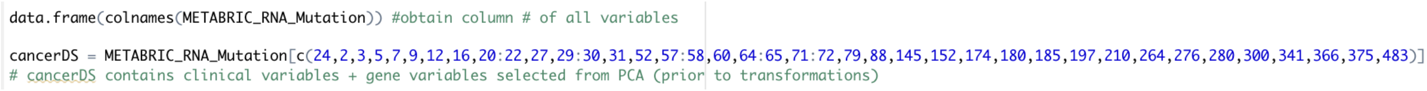
```
                        PC1   PC2   PC3   PC4   PC5
SS loadings           39.12 32.69 27.09 18.62 11.83
Proportion Var         0.08  0.07  0.06  0.04  0.02
Cumulative Var         0.08  0.15  0.20  0.24  0.26
Proportion Explained   0.30  0.25  0.21  0.14  0.09
Cumulative Proportion  0.30  0.56  0.76  0.91  1.00

Mean item complexity =  2.3
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is  0.05
 with the empirical chi square  1114030  with prob <  0

Fit based upon off diagonal values = 0.86
```

Loadings from PFA > 0.6

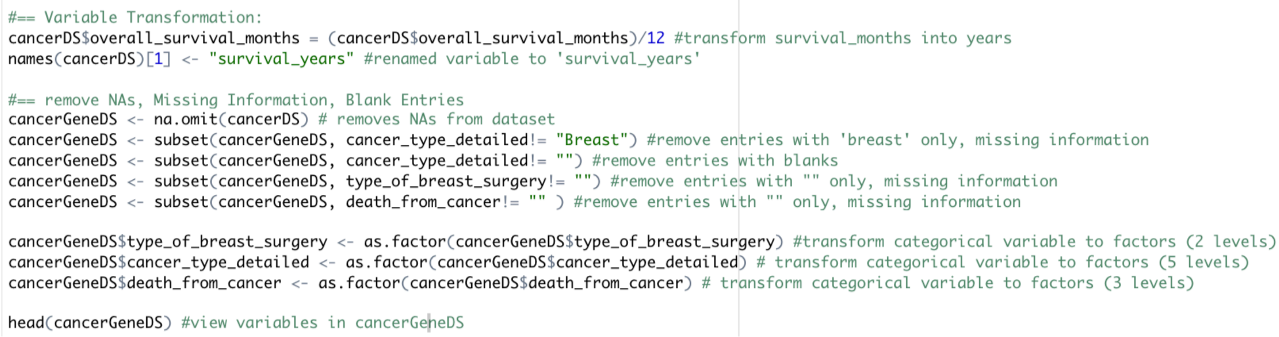
Variable	PC1	PC2	PC3	PC4
hsd17b11	0.81			
tgfbr2	0.74			
eif4e	0.70			
jak1	0.66			
itgav	0.65			
rheb	0.62			
adam10	0.60			
rb1	0.60			
ccne1		0.81		
chek1		0.77		
aurka		0.73		
e2f3		0.73		
cdc25a		0.72		
e2f2		0.71		
cdkn2a		0.61		
fancd2		0.61		
cdk1		0.61		
acvrl1			0.74	
adgra2			0.74	
dab2			0.72	
foxo1			0.69	
pdgfrb			0.66	
rps6ka2			0.61	
ccnd2			0.61	
ctcf				0.60

For cleaning the clinical variables in R, functions such as na.omit were used to omit all 'NA' entries. In addition, a copy of the original dataset was created using the selected genomic and clinical variables based on the columns of the variables. To clean the categorical clinical variables, the subset function was created to remove entries with missing text, blanks, or missing information. Categorical variables were transformed to factors using the as.factor function completing the cleaning and preparation of the data. Cleaning the dataset and condensing it allowed for a more focused analysis for the project. When working with any dataset, the data often needs to be prepared or cleaned to optimize the analysis and reporting.

R Code for Data Cleaning and Transformations:

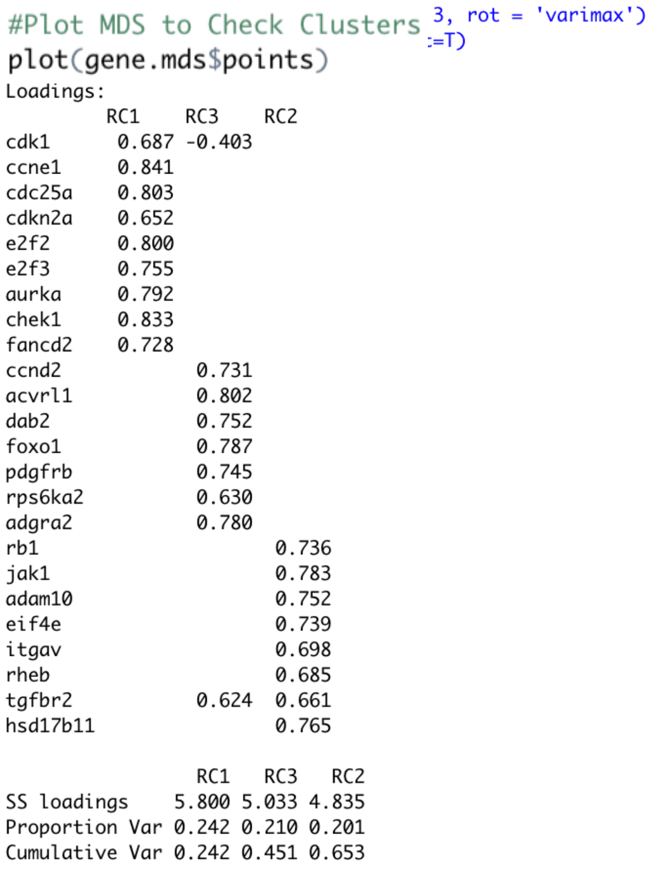


```
data.frame(colnames(METABRIC_RNA_Mutation)) #obtain column # of all variables

cancerDS = METABRIC_RNA_Mutation[c(24,2,3,5,7,9,12,16,20:22,27,29:30,31,52,57:58,60,64:65,71:72,79,88,145,152,174,180,185,197,210,264,276,280,300,341,366,375,483)]
# cancerDS contains clinical variables + gene variables selected from PCA (prior to transformations)
```

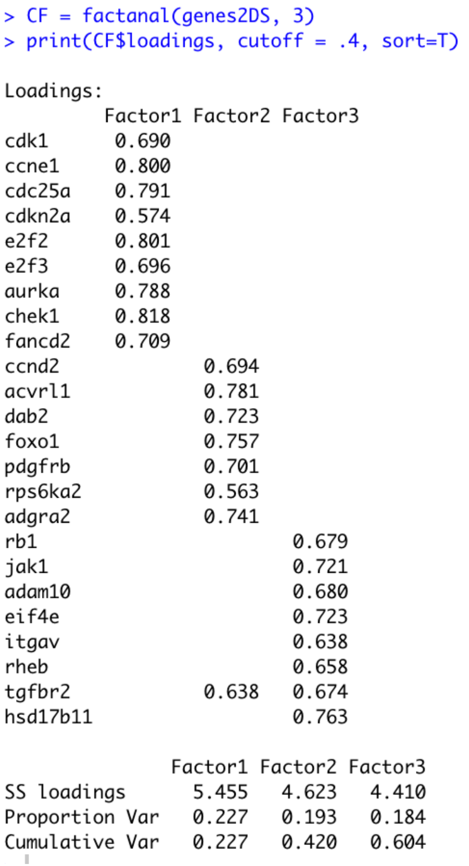


```
#== Variable Transformation:
cancerDS$overall_survival_months = (cancerDS$overall_survival_months)/12 #transform survival_months into years
names(cancerDS)[1] <- "survival_years" #renamed variable to 'survival_years'

#== remove NAs, Missing Information, Blank Entries
cancerGeneDS <- na.omit(cancerDS) # removes NAs from dataset
cancerGeneDS <- subset(cancerGeneDS, cancer_type_detailed!= "Breast") #remove entries with 'breast' only, missing information
cancerGeneDS <- subset(cancerGeneDS, cancer_type_detailed!= "") #remove entries with blanks
cancerGeneDS <- subset(cancerGeneDS, type_of_breast_surgery!= "") #remove entries with "" only, missing information
cancerGeneDS <- subset(cancerGeneDS, death_from_cancer!= "" ) #remove entries with "" only, missing information

cancerGeneDS$type_of_breast_surgery <- as.factor(cancerGeneDS$type_of_breast_surgery) #transform categorical variable to factors (2 levels)
cancerGeneDS$cancer_type_detailed <- as.factor(cancerGeneDS$cancer_type_detailed) # transform categorical variable to factors (5 levels)
cancerGeneDS$death_from_cancer <- as.factor(cancerGeneDS$death_from_cancer) # transform categorical variable to factors (3 levels)

head(cancerGeneDS) #view variables in cancerGeneDS
```

For the main analysis, Cody explored in-depth the relationship between the genomic variables by performing a full factor analysis including evaluating the correlation plots, principal component analysis, principal factor analysis and compared the results to common factor analysis. Visualizing the variables in a correlation matrix using hierarchical method, revealed three distinct groupings and one gene, *ctcf*, in its own grouping not correlated with any other variable. Performing the principal component analysis resulted in a knee at four components and the variance equals 1 criterion also confirms four components. At four components, the cumulative variance captured was 67.9%. The key for the analysis is to look at the cumulative variance at the component before and after and compare the variance captured at each component to compare and select the components with the most significance because at a certain point, the variance will simply level off. In this case, four components were sufficient, and four factors was selected for the principal factor analysis with varimax rotation. The loadings resulted in three distinct groupings and ctcf separated into its own factor. Common factor analysis was performed with four factors as a comparison and resulted in a similar grouping of factors but with ctcf and eif4e negatively related in the fourth factor. Since ctcf becomes its own factor and not strongly correlated with any other factor, it was clear from the data that the factor analysis needed to be performed again but with ctcf removed and with three factors. After removing ctcf and performing the factor analysis again, the results were similar between principal factor analysis and common factor analysis confirming the three distinct groupings for the loadings:



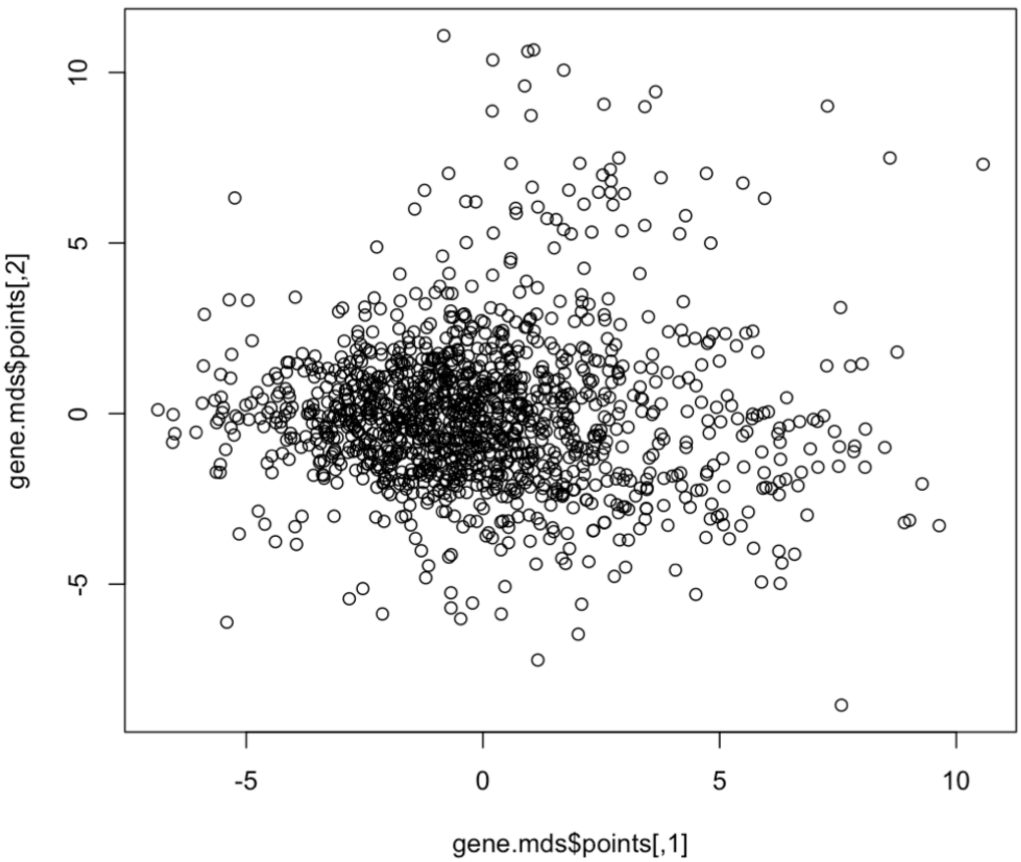
```
#Plot MDS to Check Clusters 3, rot = 'varimax')
plot(gene.mds$points)      :=T)
Loadings:
          RC1    RC3    RC2
cdk1       0.687 -0.403
ccne1      0.841
cdc25a     0.803
cdkn2a     0.652
e2f2       0.800
e2f3       0.755
aurka      0.792
chek1      0.833
fancd2     0.728
ccnd2             0.731
acvrl1            0.802
dab2              0.752
foxo1             0.787
pdgfrb            0.745
rps6ka2           0.630
adgra2            0.780
rb1                      0.736
jak1                     0.783
adam10                   0.752
eif4e                    0.739
itgav                    0.698
rheb                     0.685
tgfbr2            0.624  0.661
hsd17b11                 0.765

                 RC1   RC3   RC2
SS loadings    5.800 5.033 4.835
Proportion Var 0.242 0.210 0.201
Cumulative Var 0.242 0.451 0.653
```



```
> CF = factanal(genes2DS, 3)
> print(CF$loadings, cutoff = .4, sort=T)

Loadings:
         Factor1 Factor2 Factor3
cdk1      0.690
ccne1     0.800
cdc25a    0.791
cdkn2a    0.574
e2f2      0.801
e2f3      0.696
aurka     0.788
chek1     0.818
fancd2    0.709
ccnd2             0.694
acvrl1            0.781
dab2              0.723
foxo1             0.757
pdgfrb            0.701
rps6ka2           0.563
adgra2            0.741
rb1                       0.679
jak1                      0.721
adam10                    0.680
eif4e                     0.723
itgav                     0.638
rheb                      0.658
tgfbr2            0.638   0.674
hsd17b11                  0.763

               Factor1 Factor2 Factor3
SS loadings      5.455   4.623   4.410
Proportion Var   0.227   0.193   0.184
Cumulative Var   0.227   0.420   0.604
```

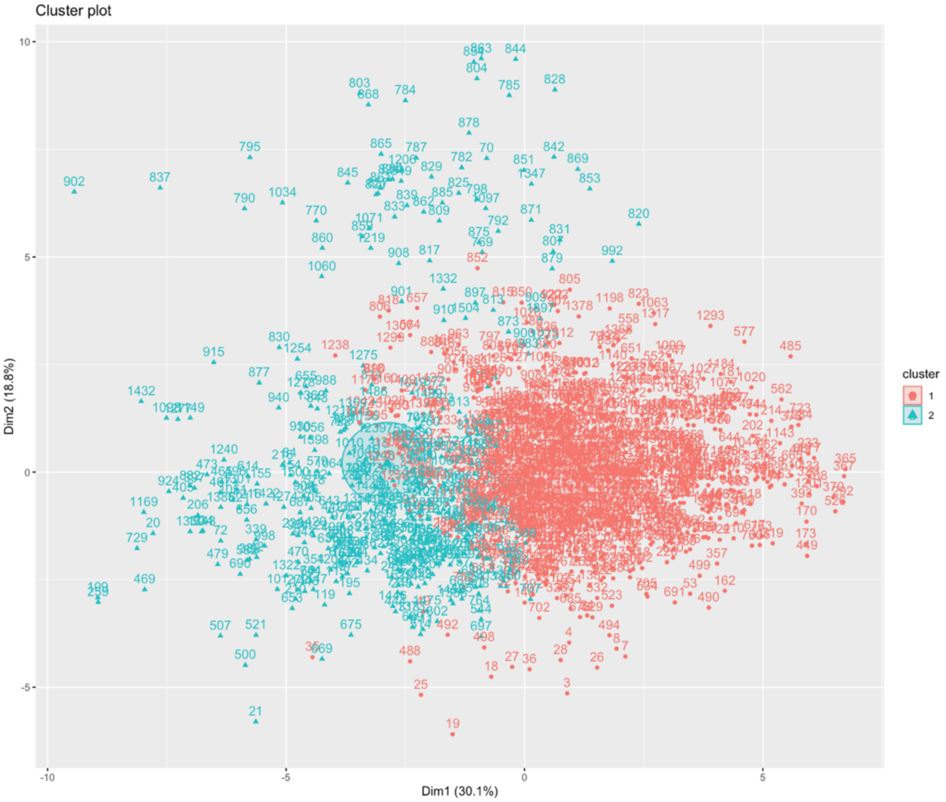
Interpreting the factors and the loadings required some domain knowledge. Cody researched the genes specifically protein coding genes, their functions, and specifically their relation to diseases and possible mutations due to cancer. Several research papers have been published specifically relating to the Cyclin Dependent Kinase (CDK) family of proteins which all concluded that the inability to regulate proteins associated with this family, directly results in cell proliferation, and unregulated cell production leads to mutations which has been documented to lead to cancer. This domain knowledge was used in determining the interpretation of the loadings, which ultimately the three factors represent genes associated with the CDK family, genes that have been documented as tumor suppressors or genes that restrict cancer progression, and lastly, genes that are associated to cell growth and regulates cell division. The last factor was the most difficult to interpret because some genes did not have documented functions and the factor included genes that when researched could cause cancer or could stop cancer progression. In summary, the genomic variables are a good fit for factor analysis and three distinct factors were found, with key attributes in terms of gene functionality. One variable, ctcf did not correlate with any other variable but this variable is directly associated with invasive breast cancer, as such, will not be removed from the regression model.

Lastly, Cody performed multidimensionality scaling on the genomic variables to determine if the data would be a good fit for factor analysis. The stress test and Shepard's diagram was evaluated. The stress resulted in a value of 0.187, which is close to 0.2, which means that the variables may not be a good fit. The Shepard's diagram also revealed no clear step line with wide distances confirming the fit for cluster analysis. Lastly, evaluating the MDS plot, showed one large dense cluster with outliers toward the top. If clusters did exist, there could potentially be two clusters.



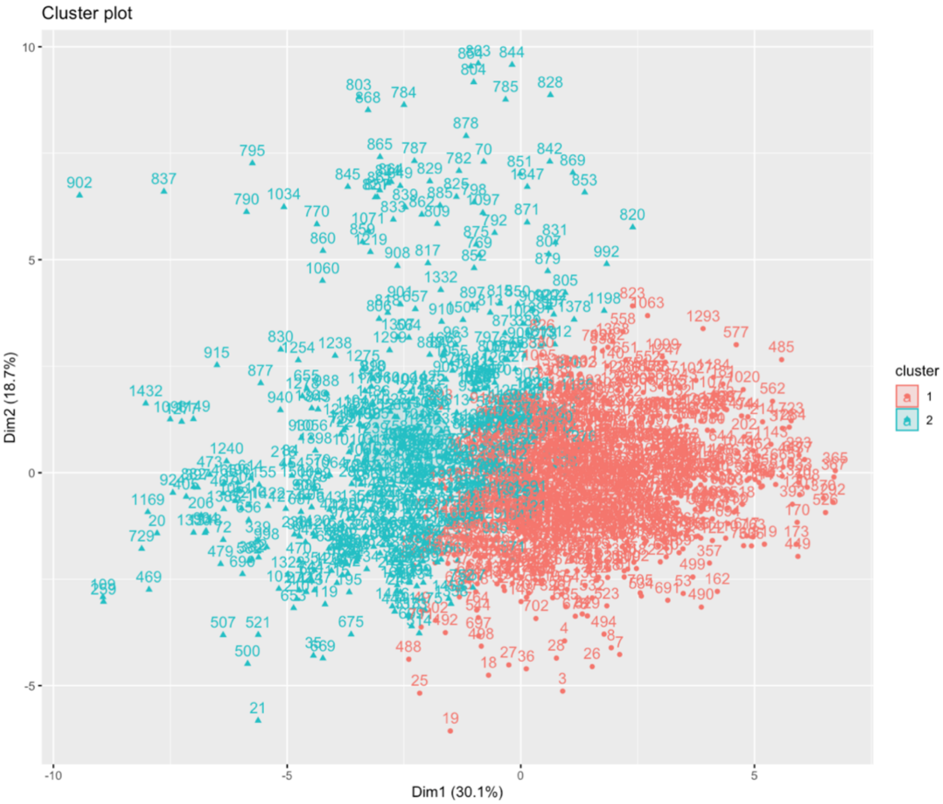
All techniques for cluster analysis including density, k-means, spectral, k-medoids and hierarchical were performed. Factoextra and cluster packages in R were used for better visualization. The average silhouette method and the gap statistic method was used to validate and optimize the number of clusters for k. This method was research and performed as part of the exploration. The average silhouette method measures the quality of the clustering and determines how well each object lies within its cluster. The gap statistic method compares the total intracluster variation for different values of k with their expected values under a distribution with no obvious clustering. Both methods resulted in k = 2, which validates that two clusters would be most optimal. The results using k-medoids and hierarchical clustering showed that distinct separated clusters could not be found. With k-medoids, the first cluster overlaps slightly less than with hierarchical clustering, but both results are very similar, and reveal that cluster analysis may not be the best fit for the data. For further exploration, k = 3 and k = 4 was also visualized and evaluated for both clustering methods. In both cases, the clusters overlapped even more, distinct lines of separation were less, and placements of classifications became more difficult. Exploring the visualization with the factoextra and cluster packages was interesting because of the different methods to view the clusters. It was easier to view the clusters and interpret the data with different method for visualization. Moreover, the factoextra and cluster packages were more advance and had more functionality than the MASS package and it was performed to visual and perform the cluster analysis using these packages. Lastly, researching about how to optimize k for cluster analysis was very useful, especially to further confirm if clustering techniques would be appropriate for the data.

```
> fviz_cluster(geneCut, data = genesDS, ellipse.type = "euclid")
```



H-Cluster Plot in Ellipse Type, Euclid, k = 2

```
> fviz_cluster(genePam, data = genesDS, ellipse.type = "euclid")
```

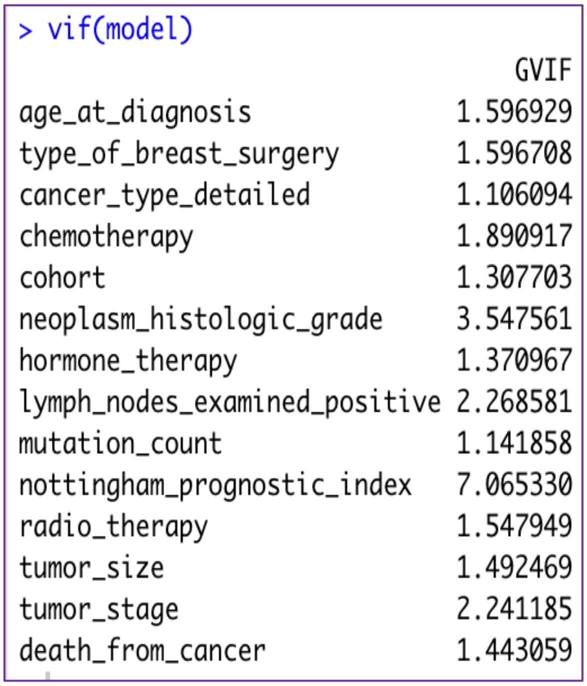


K-medoids (PAM-Cluster) in Ellipse Type, Euclid, k = 2

This project allowed for the exploration of a dataset with high dimensionality and with two distinct types of classified variables: clinical and gene expression. As this was Cody's first experience with a health and clinical related dataset, it was a rewarding experience to explore the dataset and understand the domain. From this project, we observed that dimensionality reduction through principal component analysis is very powerful and significant especially in reducing the dimensions of the data but retaining the variance. In addition, factor analysis allows for a deeper understanding of the variables and allows us to interpret the variables in a practical way, solves the issue of multicollinearity, and provides useful insights into our model and why certain variables are selected. Factor analysis also allows us to understand key variables that are significant by either being highly correlated or highly uncorrelated, both often playing significant roles in our final model. Multidimensional scaling allows us to determine if cluster analysis would be good fit for the data, which even if the scaling shows that the fit may not be great, we can still try cluster analysis as an exploration. Cluster analysis allows us to determine if data can be collected that are similar to one another but dissimilar to objects in other clusters. Clustering can then add an additional layer to understanding the variables and their relationship with each other. In using the packages in R that have advanced visualization functions for cluster analysis, visualizing the clusters and evaluating them was challenging but also a great experience. Ultimately, this project allowed for practice in key advanced analysis techniques and also provided a great opportunity to present the results in a video presentation with feedback and discussion.

Appendix H: Varsha Sajja

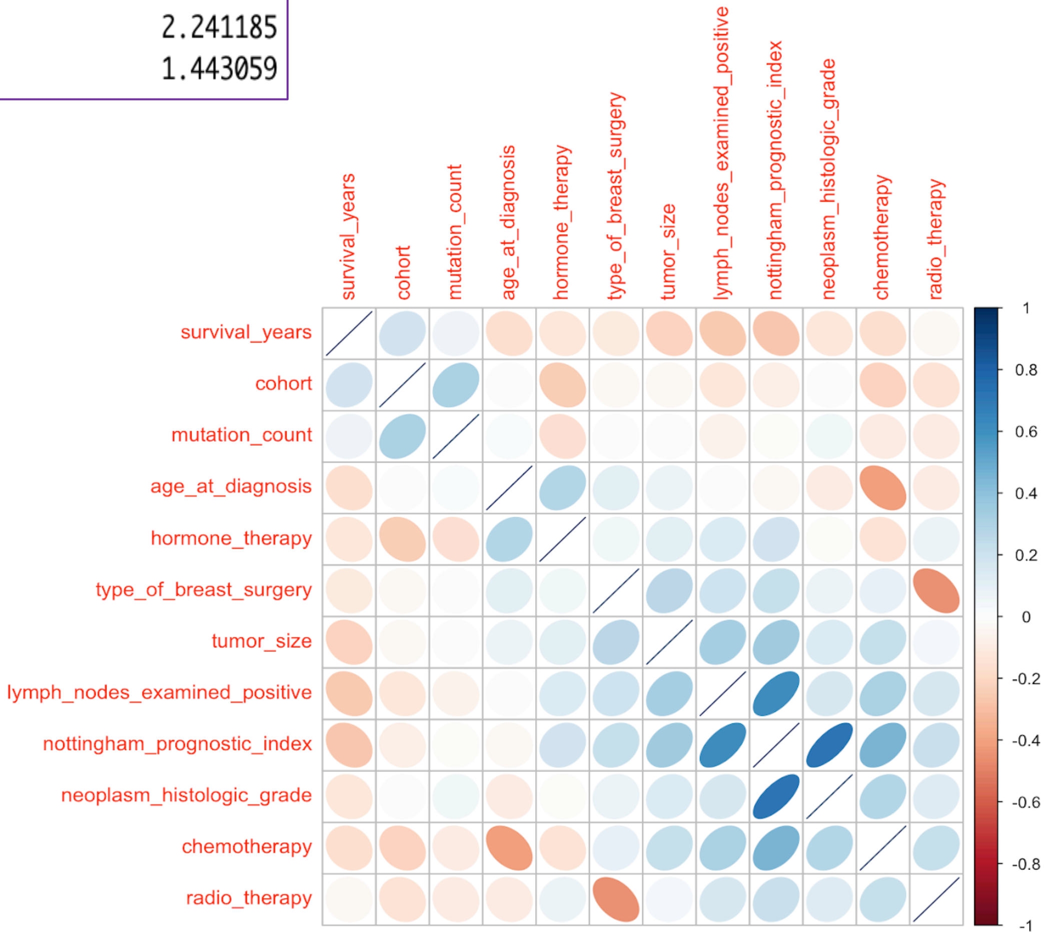
Varsha's role in this project includes exploration of all the clinical variables and their significance to any genomic data. Basic step towards checking the multicollinearity resulted in high variance inflation factor (VIF) value for *nottingham prognostic index* which is as expected because it is used for determining prognosis after the surgery for breast cancer. *Tumor size, grade of tumor and positive lymph nodes* are used for calculating prognostic index.



```
> vif(model)
                                   GVIF
age_at_diagnosis               1.596929
type_of_breast_surgery         1.596708
cancer_type_detailed           1.106094
chemotherapy                   1.890917
cohort                         1.307703
neoplasm_histologic_grade      3.547561
hormone_therapy                1.370967
lymph_nodes_examined_positive  2.268581
mutation_count                 1.141858
nottingham_prognostic_index    7.065330
radio_therapy                  1.547949
tumor_size                     1.492469
tumor_stage                    2.241185
death_from_cancer              1.443059
```

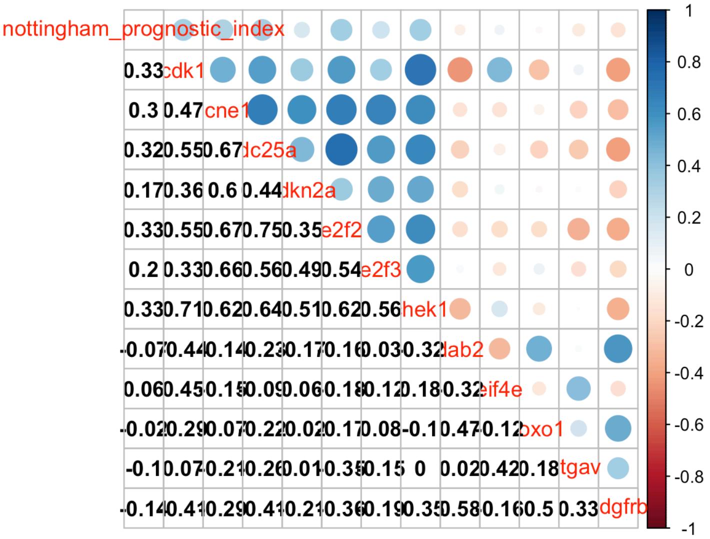
R Packages used for Clinical Data Analysis:

```
>install.packages("mosaic")
>install.packages("polycor")
>install.packages("scales")
>install.packages("gridExtra")
>install.packages("mlbench")
>install.packages("kernlab")
>install.packages("lattice")
>library(dplyr)
>library(FactoMineR)
>library(factoextra)
>library(ca)
>library(ggmosaic)
>library(vcd)
>library(polycor)
>library(MASS)
>library(mlbench)
>library(kernlab)
>library(dbscan)
>library(caret)
>library(rJava)
>library(openxlsx)
```

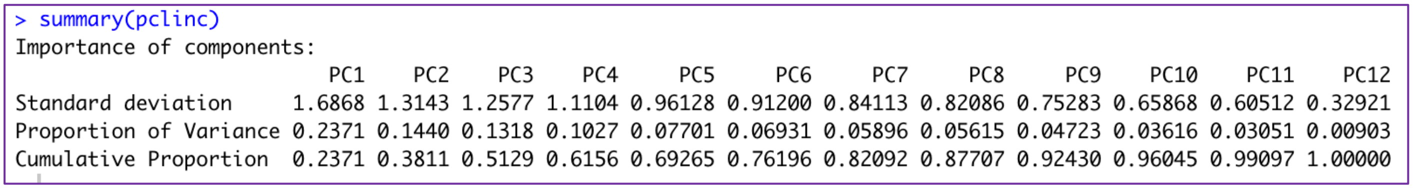


Correlation plot gives high multicollinearity around *nottingham prognostic index*. I have further explored its relation with genes which indicates few significant genes to be explored further as follows;

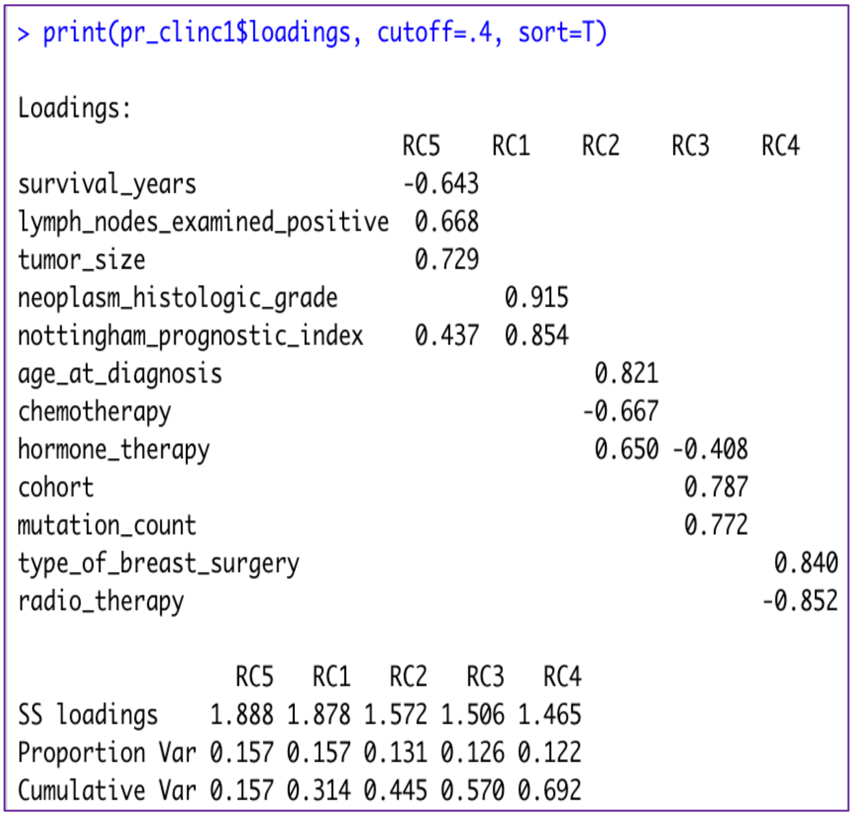
- cdk1
- ccne1
- cdc25a
- cdkn2a
- e2f2
- e2f3



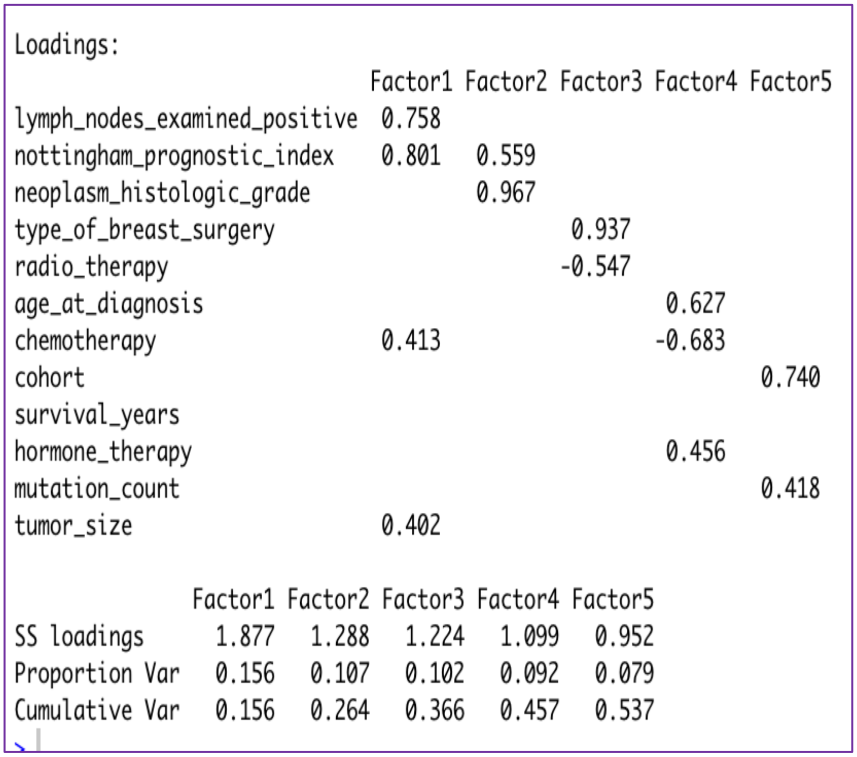
In accordance with different types of variables present in the dataset, I have performed Principal factor analysis for the principal components followed by Common factor analysis for both 4 and 5 factors of numeric data where 4 factors are taken into consideration for analysis. The principal components obtained are as below.



```
> summary(pclinc)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     1.6868 1.3143 1.2577 1.1104 0.96128 0.91200 0.84113 0.82086 0.75283 0.65868 0.60512 0.32921
Proportion of Variance 0.2371 0.1440 0.1318 0.1027 0.07701 0.06931 0.05896 0.05615 0.04723 0.03616 0.03051 0.00903
Cumulative Proportion  0.2371 0.3811 0.5129 0.6156 0.69265 0.76196 0.82092 0.87707 0.92430 0.96045 0.99097 1.00000
```



```
> print(pr_clinc1$loadings, cutoff=.4, sort=T)

Loadings:
                             RC5    RC1    RC2    RC3    RC4   
survival_years               -0.643                            
lymph_nodes_examined_positive  0.668                            
tumor_size                    0.729                            
neoplasm_histologic_grade            0.915                     
nottingham_prognostic_index   0.437  0.854                     
age_at_diagnosis                            0.821              
chemotherapy                               -0.667              
hormone_therapy                             0.650 -0.408       
cohort                                             0.787       
mutation_count                                     0.772       
type_of_breast_surgery                                    0.840
radio_therapy                                            -0.852

                 RC5   RC1   RC2   RC3   RC4
SS loadings    1.888 1.878 1.572 1.506 1.465
Proportion Var 0.157 0.157 0.131 0.126 0.122
Cumulative Var 0.157 0.314 0.445 0.570 0.692
```



```
Loadings:
                              Factor1 Factor2 Factor3 Factor4 Factor5
lymph_nodes_examined_positive  0.758
nottingham_prognostic_index    0.801   0.559
neoplasm_histologic_grade              0.967
type_of_breast_surgery                         0.937
radio_therapy                                 -0.547
age_at_diagnosis                                       0.627
chemotherapy                   0.413                  -0.683
cohort                                                         0.740
survival_years
hormone_therapy                                        0.456
mutation_count                                                 0.418
tumor_size                     0.402

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      1.877   1.288   1.224   1.099   0.952
Proportion Var   0.156   0.107   0.102   0.092   0.079
Cumulative Var   0.156   0.264   0.366   0.457   0.537
>
```

In the principal factor analysis with 4 factors, the following were interpreted:

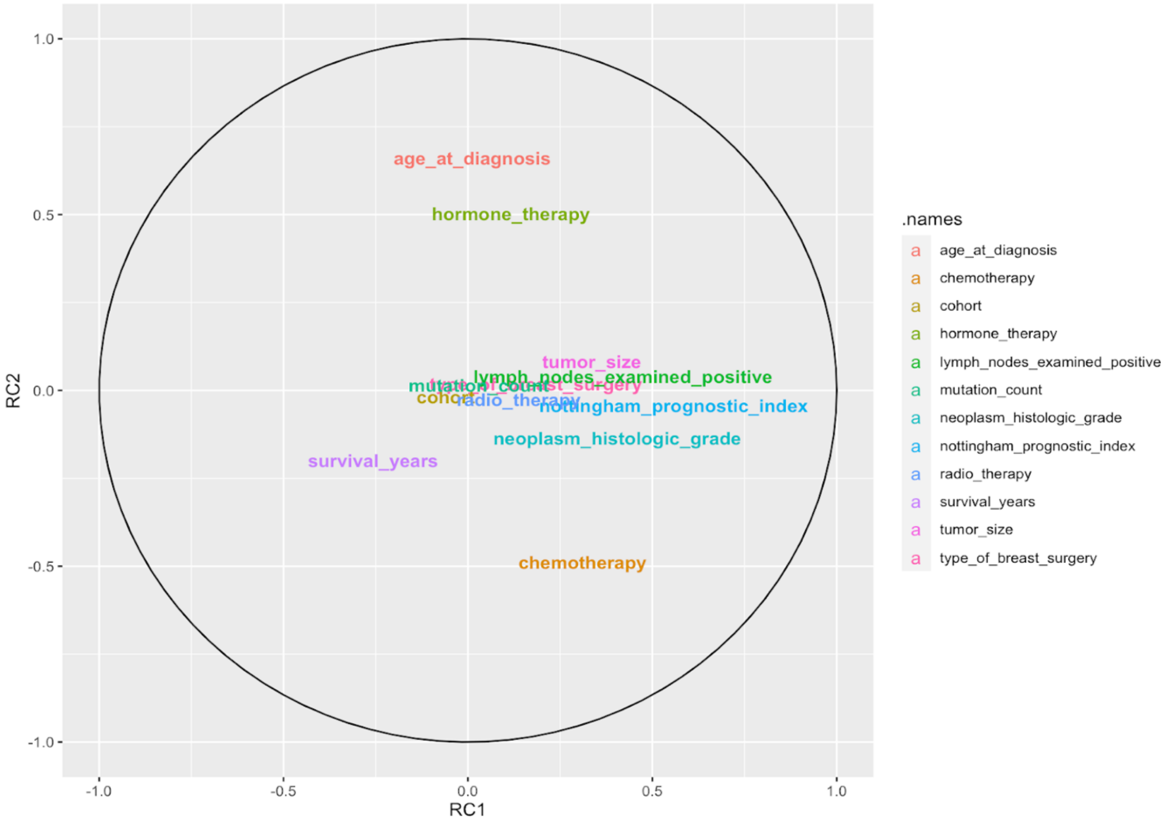
- RC1 = 0.672*neoplasm_histologic_grade* + 0.698*lymph_nodes_examined_positive* + 0.925*nottingham_prognostic_index* + 0.557*tumor_size* + 0.515*chemotherapy* - 0.426*survival_years*
  Low survival rate - The major contribution in this component is *nottingham_prognostic_index* which determines the prognosis following breast cancer surgery. Thus, other components such as *neoplasm_histologic_grade*, *lymph_nodes_examined_positive*, *tumor_size* are also having contributions. Also noticed that *chemotherapy* has a significant contribution. *Survival*_years is negatively correlated which illustrates that tumor has been invaded and chemotherapy doesn't work.

- RC2 = 0.838*age_at_diagnosis* - 0.619*chemotherapy* + 0.638*hormone_therapy*
  Hormone Therapy - The major contribution in this component is *age*_at_*diagnosis* which determines the patient's age at the time of prognosis. Chemotherapy is negatively correlated here whereas *hormone_therapy* works better.

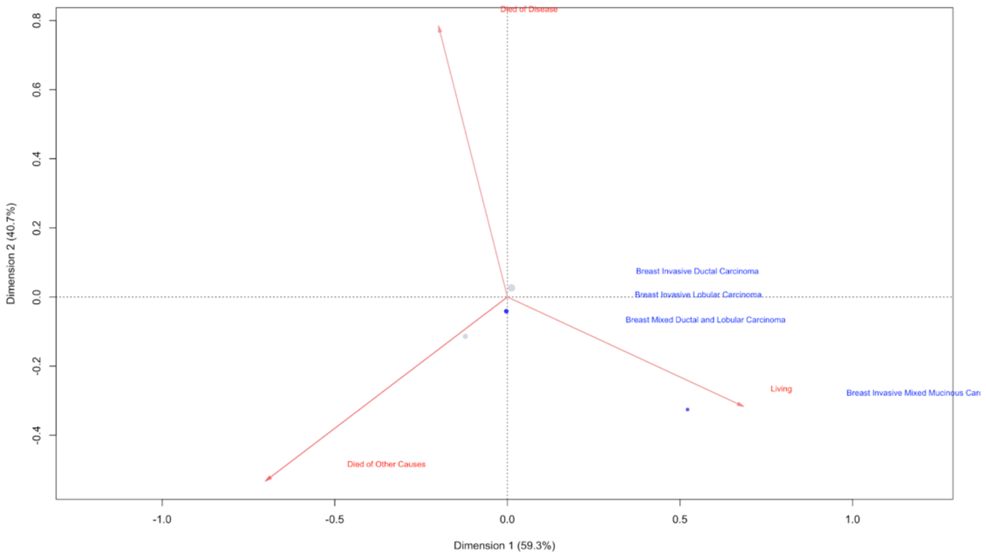
- RC3 = 0.787*cohort* + 0.739*mutation_count*
  Cancer Characteristics - Both *cohort* and *mutation_count* are highly contributed towards this component which determines the gene variables having shared characteristics of relevant mutations.

- RC4 = 0.829*radio_therapy* - 0.832*type_of_breast_surgery*
  Surgery Vs. Therapy - *type_of_breast_surgery* is negatively correlated in this component and *radio_therapy* has high positive correlation with others.

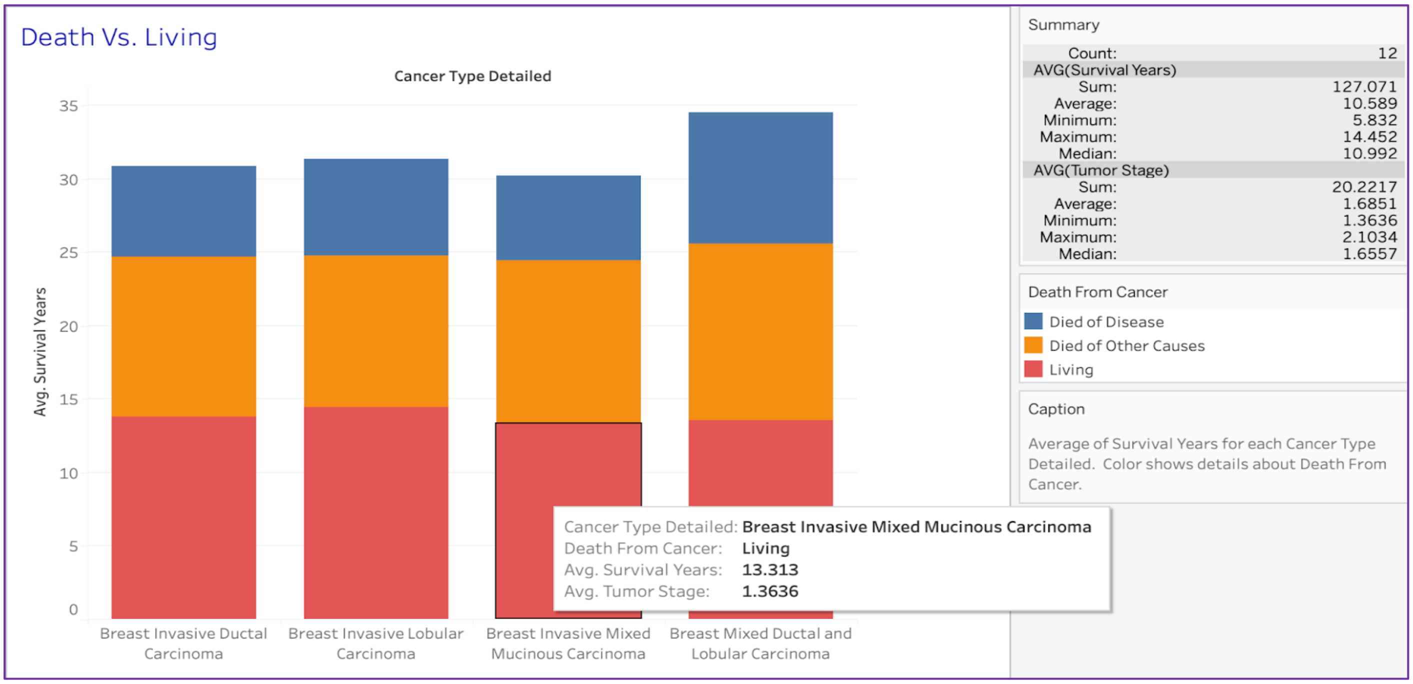
Finally, the PFA and CFA have also been performed with 5 factors resulting in 69.2% variance in data. Although the variance captured is greater than in 4 factors, it doesn't make a difference that 4 factors are sufficient for analysis. We can conclude here that the *survival years* is negatively correlated here as well. CFA gives an interpretation of forming a factor with *cohort* and *mutation count* separately. By taking 4 sufficient factors into consideration, VARIMAX rotation has been performed and the majority of significant clinical variables are around the origin.



Correspondence Analysis is performed for categorical Variables of clinical data. The 2D representation gives us an idea on the direction of arrows (eigen vectors) denoting the importance.



The below visualization gives an understanding on the *cancer type detailed* and *survival years.*



Death Vs. Living

Cancer Type Detailed
Avg. Survival Years
35
30
25
20
15
10
5
0
Breast Invasive Ductal Carcinoma
Breast Invasive Lobular Carcinoma
Breast Invasive Mixed Mucinous Carcinoma
Breast Mixed Ductal and Lobular Carcinoma

Cancer Type Detailed: **Breast Invasive Mixed Mucinous Carcinoma**
Death From Cancer: **Living**
Avg. Survival Years: **13.313**
Avg. Tumor Stage: **1.3636**

Summary

Count:	12
AVG(Survival Years)	
Sum:	127.071
Average:	10.589
Minimum:	5.832
Maximum:	14.452
Median:	10.992
AVG(Tumor Stage)	
Sum:	20.2217
Average:	1.6851
Minimum:	1.3636
Maximum:	2.1034
Median:	1.6557

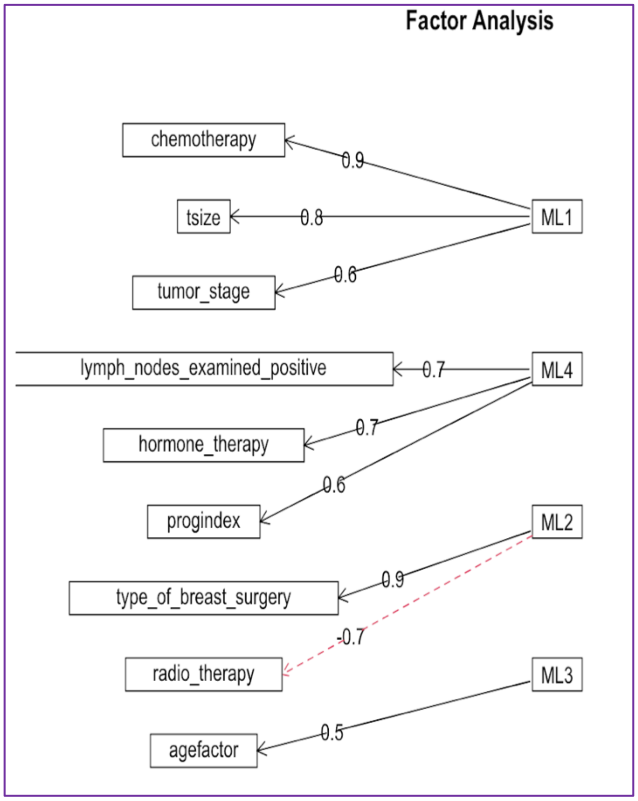
Death From Cancer

🟦 Died of Disease
🟧 Died of Other Causes
🟥 Living

Caption

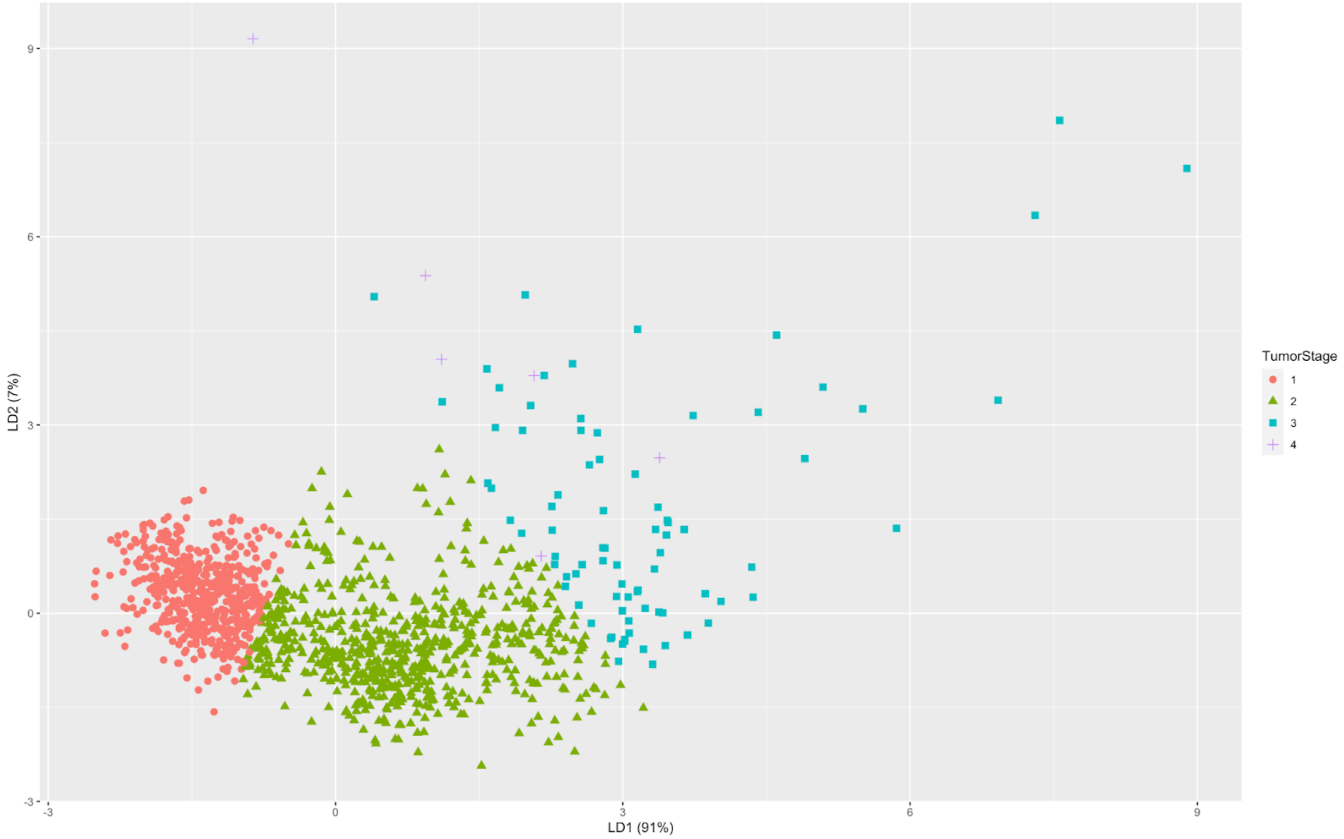
Average of Survival Years for each Cancer Type Detailed. Color shows details about Death From Cancer.

Polychoric factor analysis concludes the ordinal variable tumor stage has its significance on other clinical data. The factor rotation has been applied to the analysis.



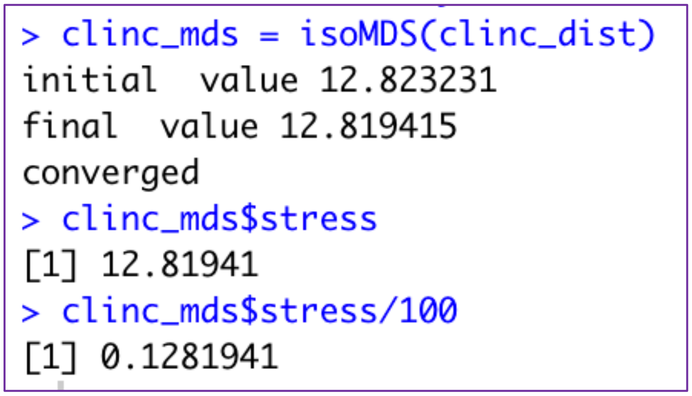
Linear Discriminant Analysis

This dimensionality reduction step is used as a pre-processing step for pattern-classification. In linear discriminant analysis, linear discriminants are obtained which are later projected in 2 dimensions. Here, tumor stage is considered as a factor in considering the analysis. We get the 4 tumor stages distinguished clearly. Also, few of the outliers are detected in tumor stage 3.



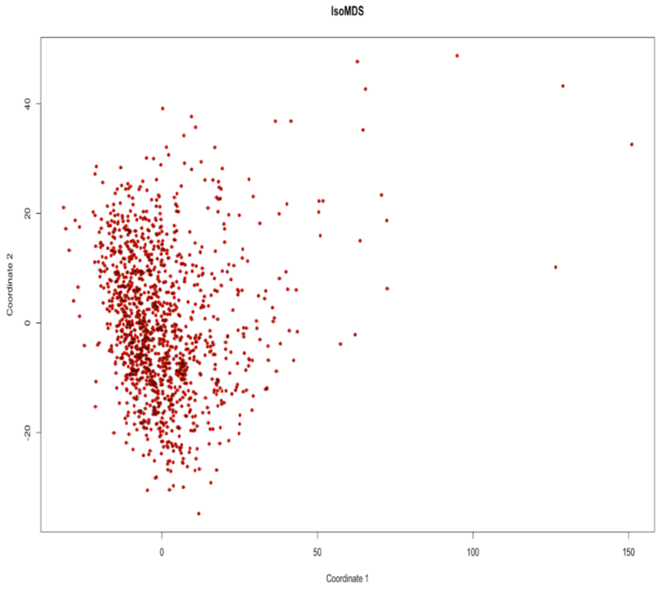
Cluster Analysis

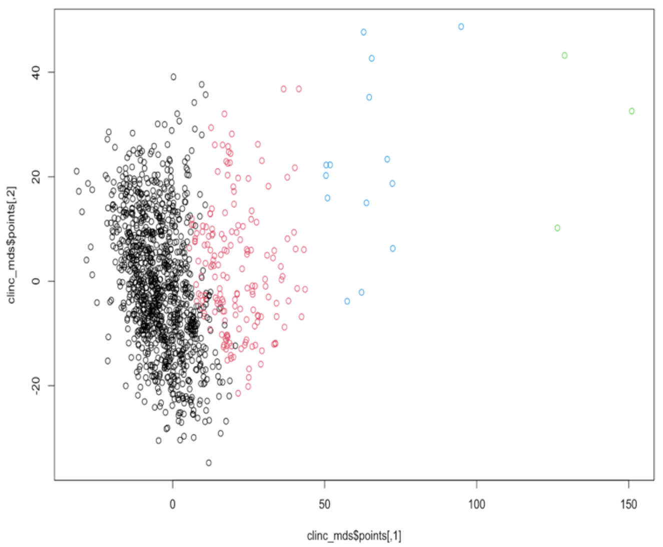
Multidimensional Scaling is performed to determine the clusters in data. "isoMDS" function is extensively used with distance to calculate stress. Stress is reported to be 0.1281 which is between 0 and 1 suggesting a good fit.



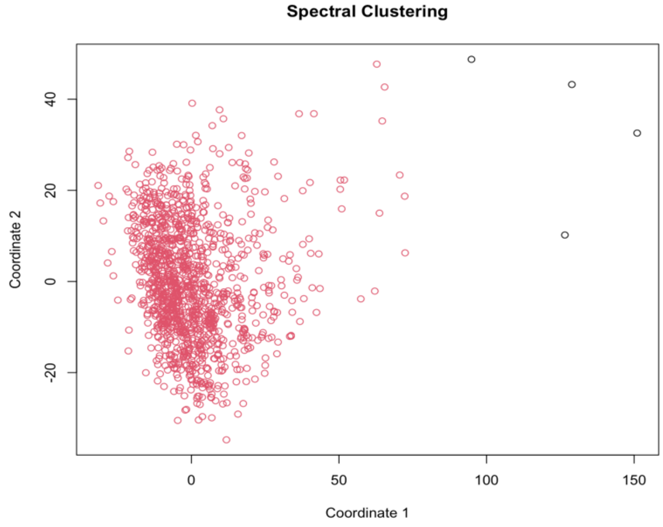
```
> clinc_mds = isoMDS(clinc_dist)
initial  value 12.823231
final  value 12.819415
converged
> clinc_mds$stress
[1] 12.81941
> clinc_mds$stress/100
[1] 0.1281941
```

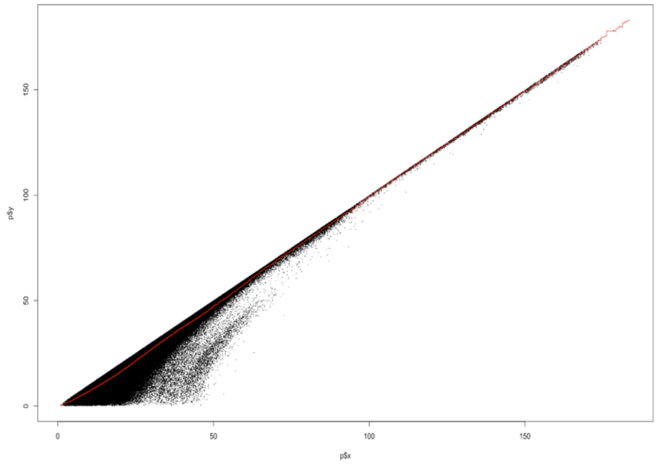
The plot for MDS concentrates over one cluster with few outliers. On vaguely performing hierarchical clustering, the plot for MDS is converted with 4 clusters leaving few outliers.



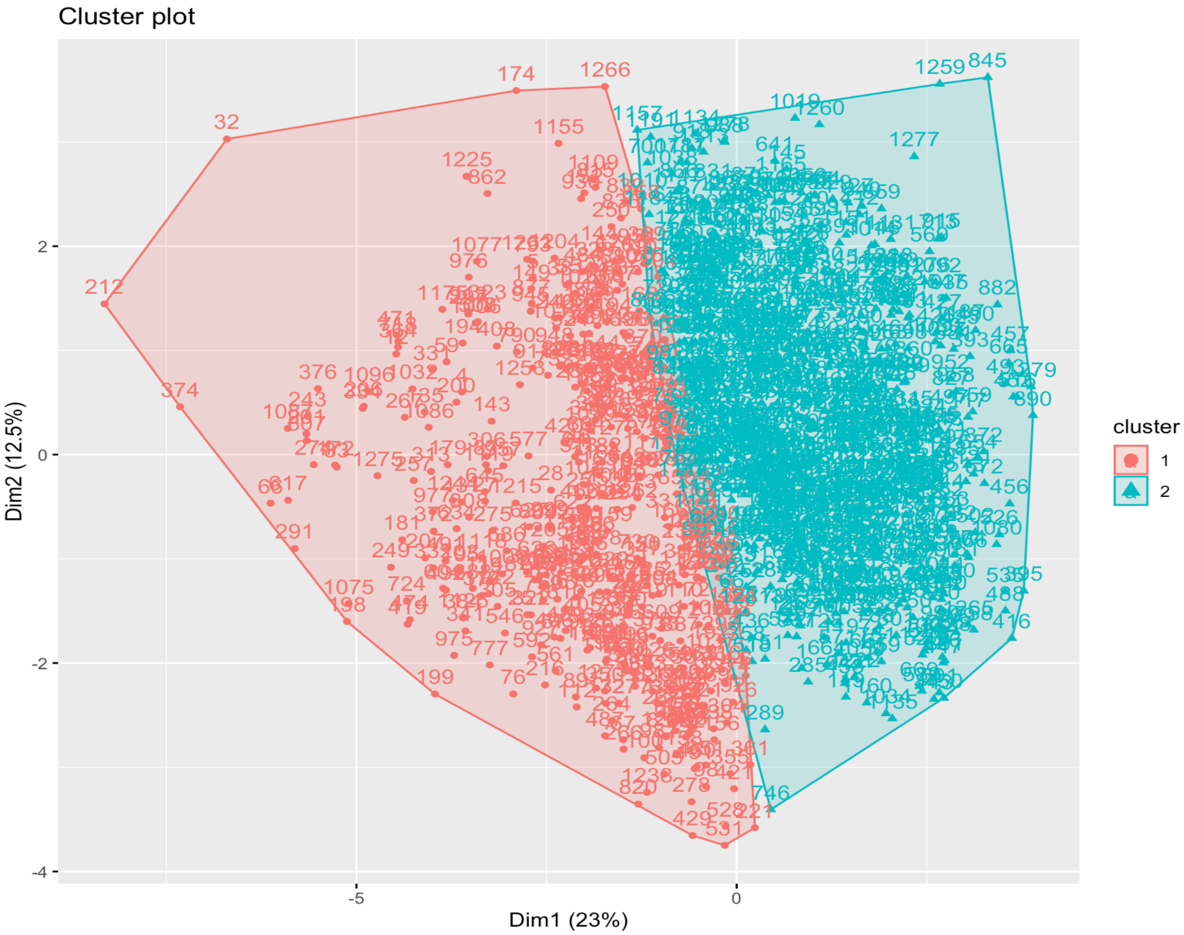


Another clustering technique called Spectral clustering is performed, which distinguishes the outliers by color. Although Shepard's method is used to distinguish the data which has big initial deviations, but outliers have been drastically reduced at the end forming the data around the red line.





Now k-means clustering has been performed on the data for clusters. We get two clusters formed up separating a line between them. This indicates that the clinical data can be distinguished into two clusters.



These findings have been used for regression techniques performed on the data considering the significance of variables and classification performed which also settles *tumor stage* to be an important factor for its correlation with other variables. In general, the stage of tumor utmost decides the cancer level and cure for it.

Although the goals of this project lead me to explore the clinical representation of breast cancer and ways to treat it with diagnosis, I took an initiative in learning how the genomic variables play a key role in the invasion of tumors in breast cancer patients as well. With the available clinical data in the breast cancer dataset, I have analyzed how tumor stage affects the overall survival years and also illustrated how independent therapies such as chemotherapy, hormone therapy and radio therapy aids in treating the patients depending on their age at prognosis. The model we worked on also consists of the same significant clinical variables. During the analysis, cancer type also had an impact on death of disease where the common type of breast cancer has been identified. Applying the major techniques learned through the course work in this project, gave me complete understanding towards the real-world data and its relationship with significant terms. The clustering analysis overlooked the tumor stages in my model and distinguished the stages of tumor in two dimensions. Nevertheless, breast cancer is more likely to be metastatic (spreads to other parts of the body). If the source dataset had more information on such data, then the analysis of survival rate could be precise. We believe that prevention is better than cure, but tumor cells need to be detected prior to treatment.

Appendix I: Aaron Gregory

Aaron's role was to create a complete regularized regression model that could be used to predict survival years for breast cancer patients. At the outset of the project, he looked at different corr plots and pair panels to try and find interaction variables, data transformations and interesting dependent variables to investigate. Based on our dataset it was intuitive to look at survival and survival length as these variables would be particularly important to patients recently diagnosed with breast cancer. Analyzing the pair panel, we can see that the largest correlations for survival years is tumor stage and Nottingham prognostic index. Both Nottingham prognostic index and tumor stage were prominent variables throughout our analysis and were confirmed to be important in our CFA/PFA analysis. This early analysis gave us some insight into what our final model would look like as there are not many highly correlated variables in our data set.

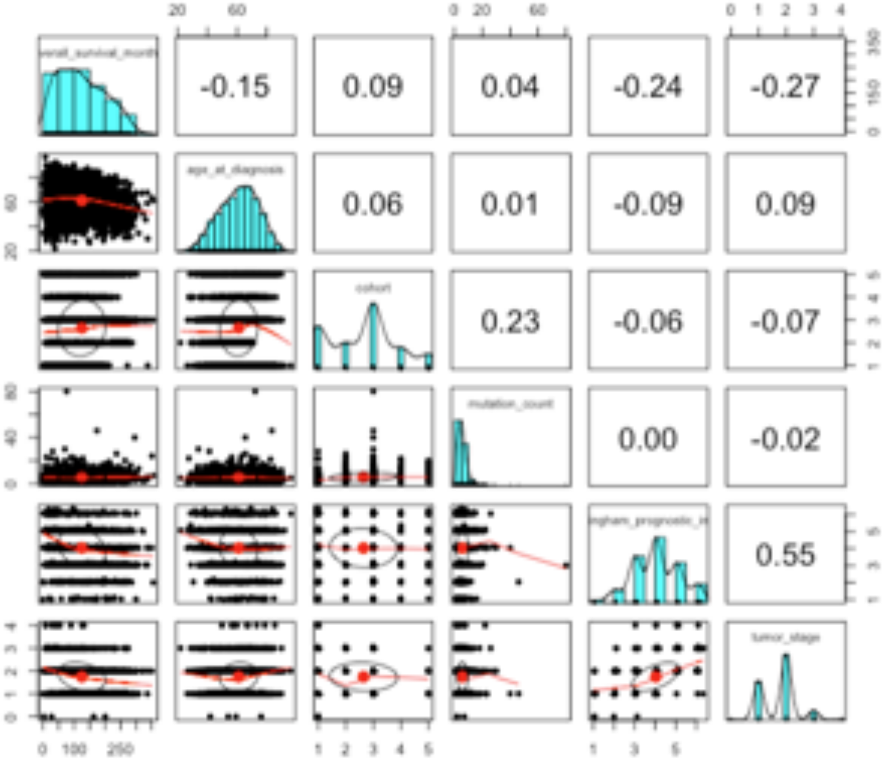


Figure 1: Pairs Panel of most important clinical data

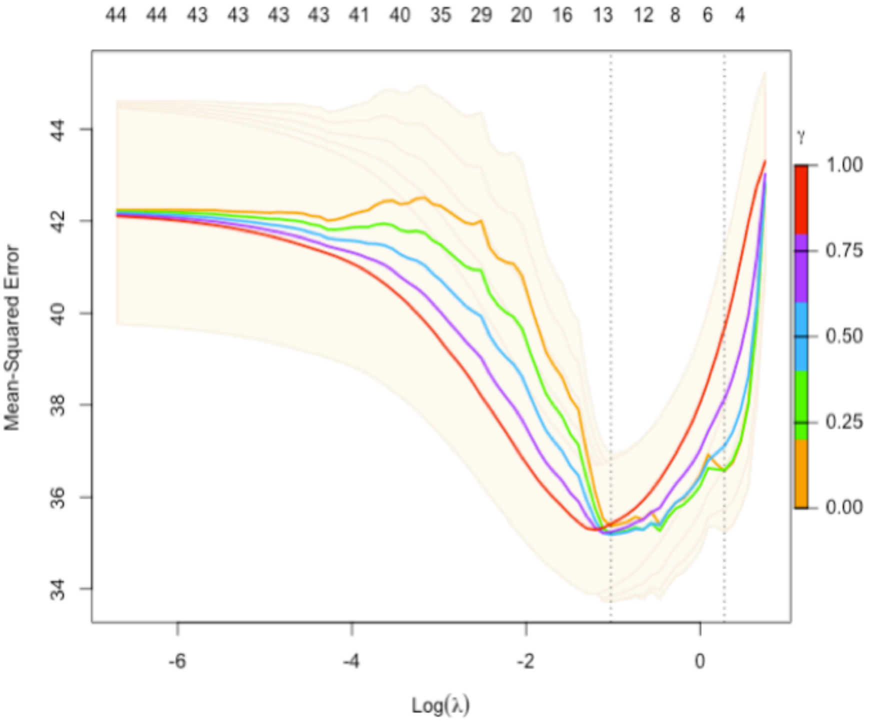
One transformation we chose was turning breast cancer type into a binary variable so that it could be used in the regression model. Based on our initial analysis we came up with several interaction variables that we thought would be useful to explore based on the correlations and hypothesis about cancer growth. These variables include:

- Radiotherapy and tumor size
- Radiotherapy and tumor stage
- Chemotherapy and tumor size
- Chemotherapy and tumor stage
- Hormone therapy and tumor size
- Hormone therapy and tumor stage

We tried several different model building processes in order to find the best possible model for survival years including, backwards step regression, all-subsets regression, ridge regression, lasso regression and relaxed lasso regression. In the end we found that a relaxed lasso gave us the best model as it provides variable selection without depressing the beta values too greatly.

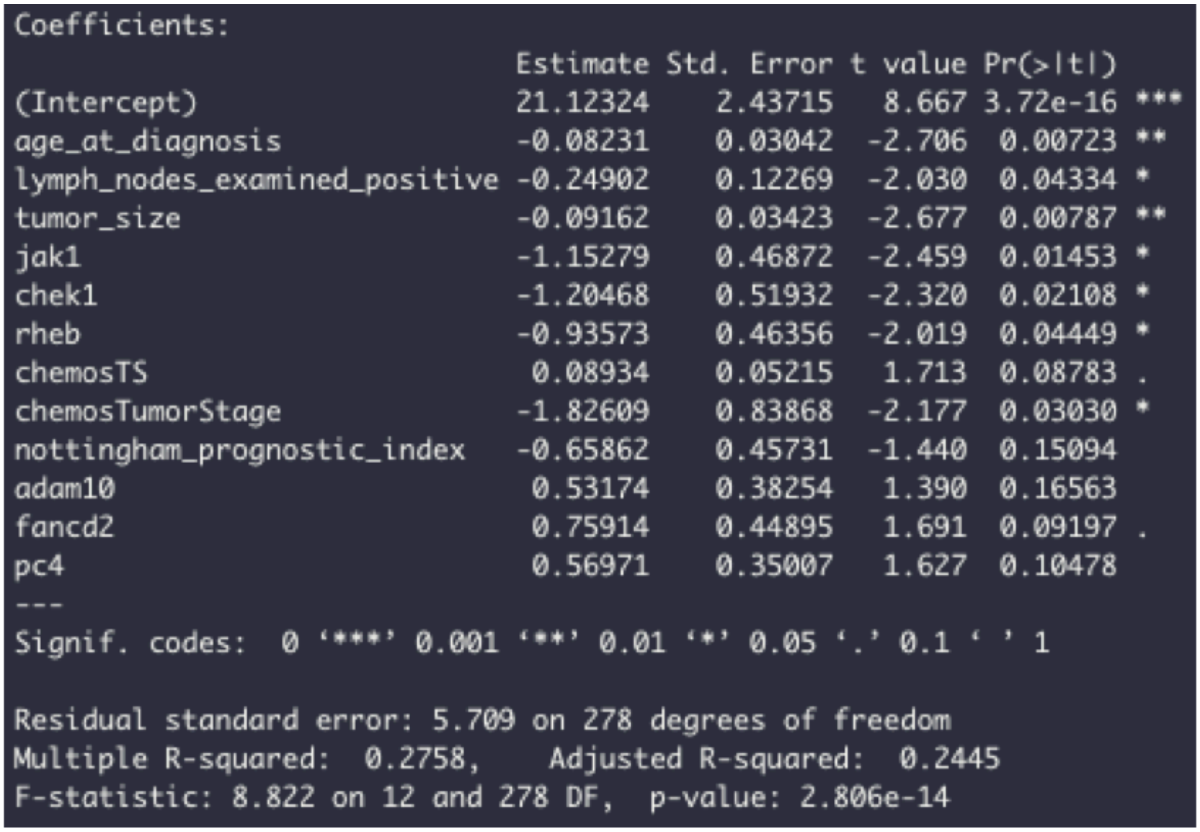
Looking at the most basic three regression models, (full model, backwards stepwise regression and all subsets) we see a low adjusted R^2 value along with a sizable increase in the RMSE from the training set to the test set. This shows that we are overfitting our data and could have some multicollinearity present in our models. The basic models were not very effective as they did a poor job with variable selection and reducing multicollinearity.

A summary of the regression models we looked at are included in the technical report. Ridge and Lasso techniques improved our models from the base model but in the end a relaxed lasso approach was the most effective model. What we found is that a lot of variables have a small impact on our dependent variable. When manually building the model, we found that different variables could be substituted to generate equivalently efficient models. For example, if RHEB was removed and replaced with PCA2 the difference in RMSE and adjusted R^2 value was negligible. Something else that we found was that when using relaxed lasso our final model resulted in a very low lambda value (0.25) as seen in the graph to the right. This shows that a lot of the multicollinearity variables were selected out of model and there wasn't much regularization needed on the final set of variables.



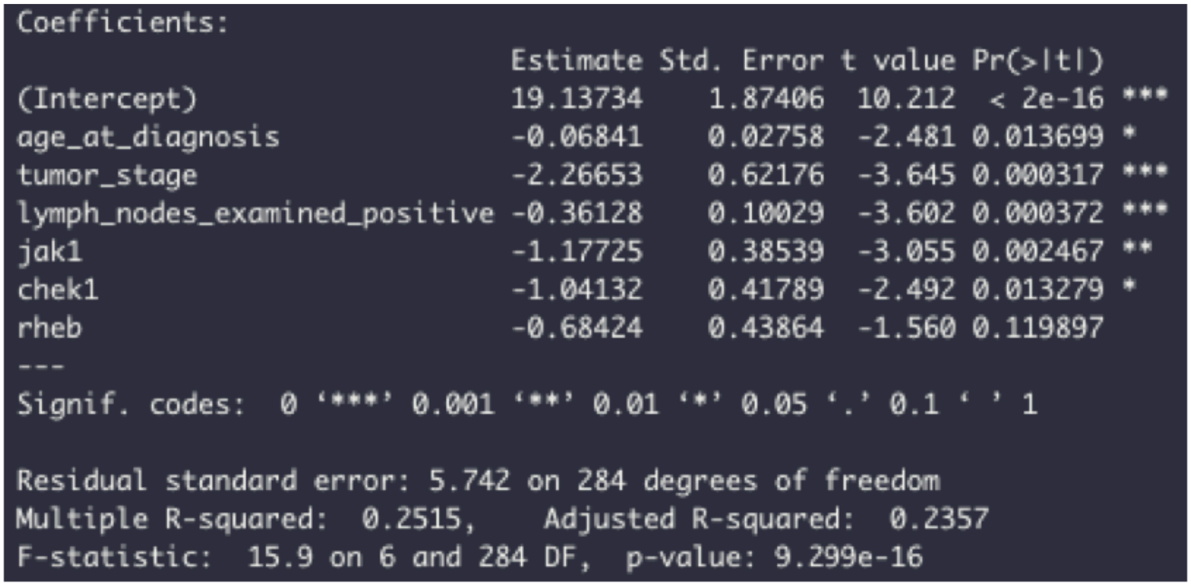
The main takeaway from this portion of the project was how complicated it can be to model a complex issue. In class we have frequently used simple data sets with more or less clear-cut relationships and results. When analyzing our data set and trying to produce a model it really showed how in a real-world situation having all of the variables and data you would want is not always possible. The issues that our final model has exemplifies the complexity of our bodies and health in general. There is no dominating factor or gene that will explain a particular point of interest. Our bodies and health are controlled by an ecosystem of factors that are hard to replicate in a simple model.

Model Results



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  21.12324    2.43715   8.667 3.72e-16 ***
age_at_diagnosis             -0.08231    0.03042  -2.706  0.00723 ** 
lymph_nodes_examined_positive -0.24902    0.12269  -2.030  0.04334 *  
tumor_size                   -0.09162    0.03423  -2.677  0.00787 ** 
jak1                         -1.15279    0.46872  -2.459  0.01453 *  
chek1                        -1.20468    0.51932  -2.320  0.02108 *  
rheb                         -0.93573    0.46356  -2.019  0.04449 *  
chemosTS                      0.08934    0.05215   1.713  0.08783 .  
chemosTumorStage             -1.82609    0.83868  -2.177  0.03030 *  
nottingham_prognostic_index  -0.65862    0.45731  -1.440  0.15094    
adam10                        0.53174    0.38254   1.390  0.16563    
fancd2                        0.75914    0.44895   1.691  0.09197 .  
pc4                           0.56971    0.35007   1.627  0.10478    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.709 on 278 degrees of freedom
Multiple R-squared:  0.2758,	Adjusted R-squared:  0.2445 
F-statistic: 8.822 on 12 and 278 DF,  p-value: 2.806e-14
```

Backwards Stepwise Regression



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  19.13734    1.87406  10.212  < 2e-16 ***
age_at_diagnosis             -0.06841    0.02758  -2.481 0.013699 *  
tumor_stage                  -2.26653    0.62176  -3.645 0.000317 ***
lymph_nodes_examined_positive -0.36128    0.10029  -3.602 0.000372 ***
jak1                         -1.17725    0.38539  -3.055 0.002467 ** 
chek1                        -1.04132    0.41789  -2.492 0.013279 *  
rheb                         -0.68424    0.43864  -1.560 0.119897    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.742 on 284 degrees of freedom
Multiple R-squared:  0.2515,	Adjusted R-squared:  0.2357 
F-statistic:  15.9 on 6 and 284 DF,  p-value: 9.299e-16
```

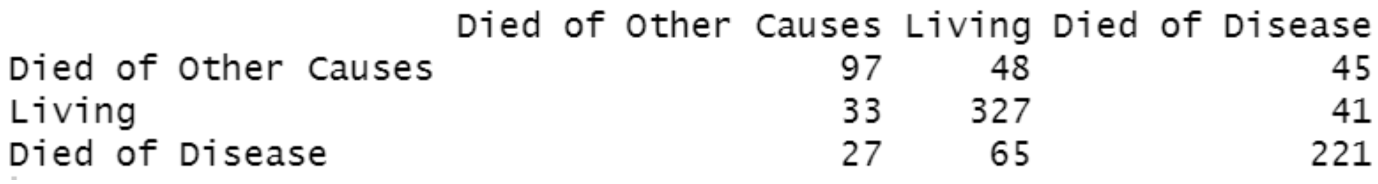
All Subsets Regression

Appendix J: Evan Morton

Evan’s contribution to the project focused on exploring multinomial logistic regression for the death from cancer variable. The model building was done by using the *nnet* package and the *multinom* function. The code below is shown for all variable models.

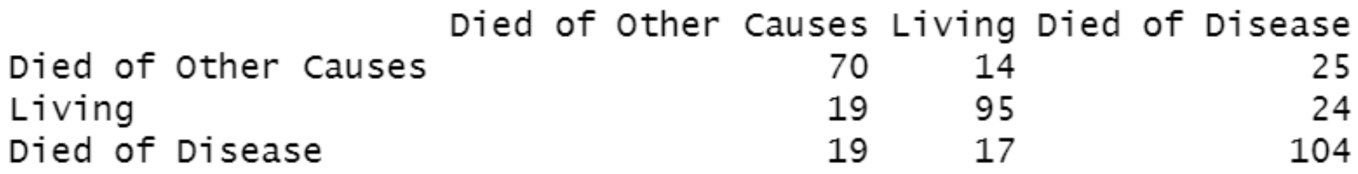
multi_mod <- multinom(death_from_cancer ~ ., data = CanGeneD_train)

Below is a confusion matrix that shows the predicted values versus the actual values of the model. The model was created using a training dataset that was split from the original data. This way we can test the model’s performance on the test set as well. The confusion matrix is for the training set. As shown, it is not perfect, but is able to correctly categorize a large portion of the data. This model as shown has an accuracy rate of 71.38%.



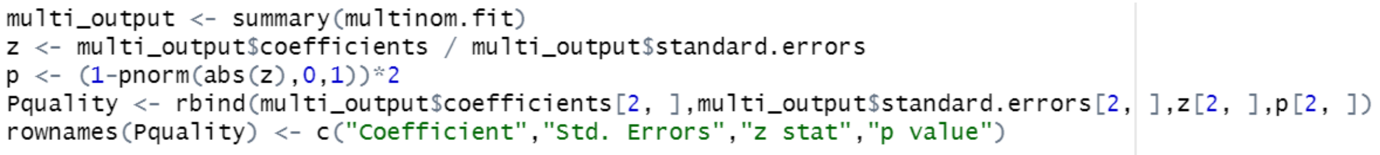
```
                   Died of Other Causes Living Died of Disease
Died of Other Causes                 97     48              45
Living                               33    327              41
Died of Disease                      27     65             221
```

To test the model further, it was applied to the test dataset and its performance was compared. The confusion matrix for the model applied to the test set is shown below. Like the training data, there is a large portion categorized correctly but many entries are incorrect. It has an accuracy of 69.9%. This is only a drop of 1.5%, which is good because it does not immediately indicate that there is overfitting.

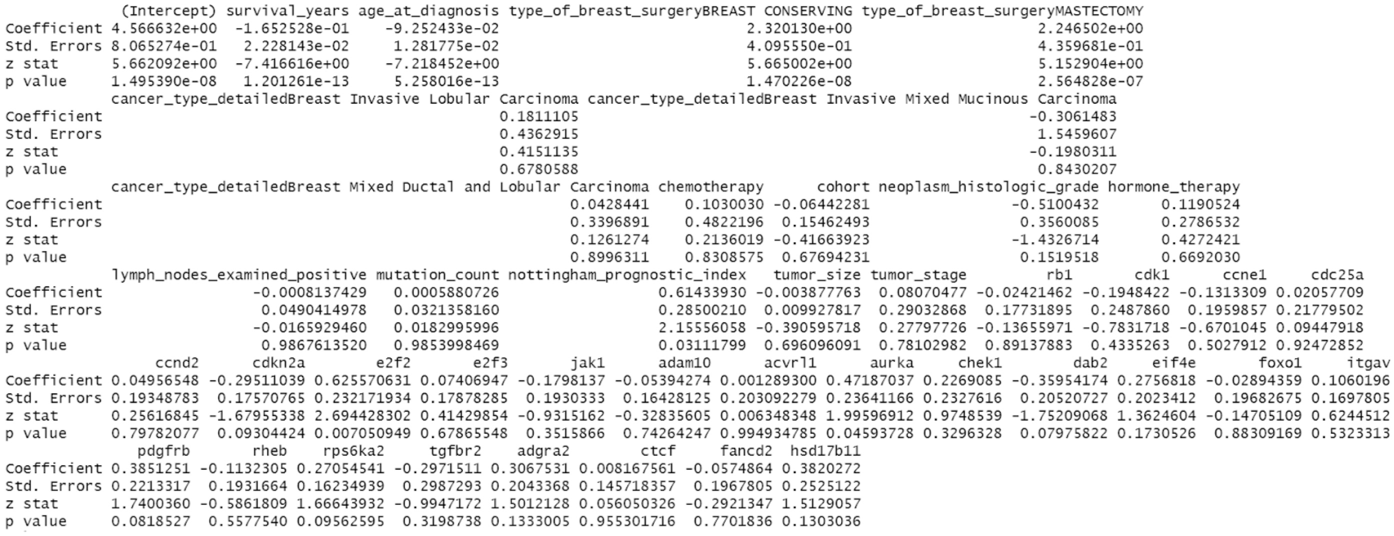


```
                   Died of Other Causes Living Died of Disease
Died of Other Causes                 70     14              25
Living                               19     95              24
Died of Disease                      19     17             104
```

Shown below is a generated table that gives the coefficients, standard error, z-score, and p-value. The code to create that table is shown above it. From this table, it is clear that not all of the variables are needed in the model. Based on the p-values, I could see that survival_years, age_at_diagnosis, type_of_breast_surgery, nottingham_prognostic_index, e2f2, and aurka were the variables that were statistically significant.

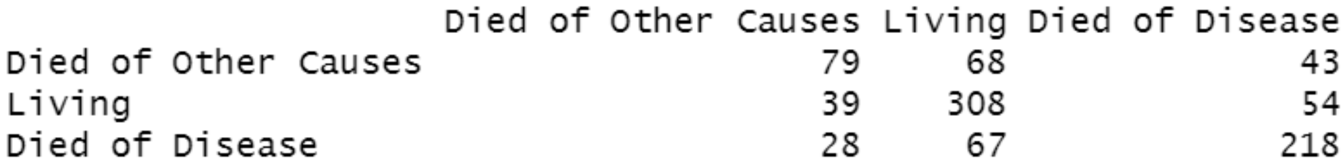


```
multi_output <- summary(multinom.fit)
z <- multi_output$coefficients / multi_output$standard.errors
p <- (1-pnorm(abs(z),0,1))*2
Pquality <- rbind(multi_output$coefficients[2, ],multi_output$standard.errors[2, ],z[2, ],p[2, ])
rownames(Pquality) <- c("Coefficient","Std. Errors","z stat","p value")
```

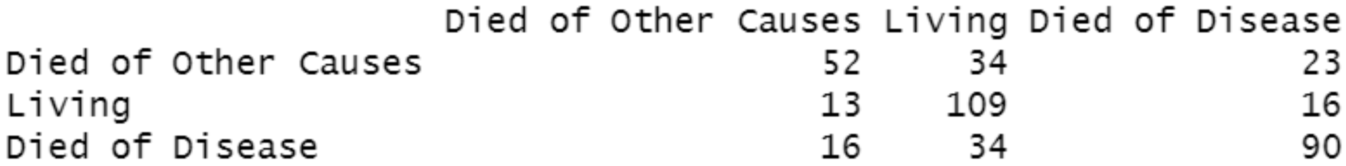


```
            (Intercept) survival_years age_at_diagnosis type_of_breast_surgeryBREAST CONSERVING type_of_breast_surgeryMASTECTOMY
Coefficient 4.566632e+00  -1.652528e-01    -9.252433e-02                            2.320130e+00                     2.246502e+00
Std. Errors 8.065274e-01   2.228143e-02     1.281775e-02                            4.095550e-01                     4.359681e-01
z stat      5.662092e+00  -7.416616e+00    -7.218452e+00                            5.665002e+00                     5.152904e+00
p value     1.495390e-08   1.201261e-13     5.258016e-13                            1.470226e-08                     2.564828e-07
            cancer_type_detailedBreast Invasive Lobular Carcinoma cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma
Coefficient                                             0.1811105                                                   -0.3061483
Std. Errors                                             0.4362915                                                    1.5459607
z stat                                                  0.4151135                                                   -0.1980311
p value                                                 0.6780588                                                    0.8430207
            cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma chemotherapy      cohort neoplasm_histologic_grade hormone_therapy
Coefficient                                                     0.0428441    0.1030030 -0.06442281                -0.5100432       0.1190524
Std. Errors                                                     0.3396891    0.4822196  0.15462493                 0.3560085       0.2786532
z stat                                                          0.1261274    0.2136019 -0.41663923                -1.4326714       0.4272421
p value                                                         0.8996311    0.8308575  0.67694231                 0.1519518       0.6692030
            lymph_nodes_examined_positive mutation_count nottingham_prognostic_index   tumor_size tumor_stage         rb1       cdk1      ccne1     cdc25a
Coefficient                 -0.0008137429   0.0005880726                  0.61433930 -0.003877763  0.08070477 -0.02421462 -0.1948422 -0.1313309 0.02057709
Std. Errors                  0.0490414978   0.0321358160                  0.28500210  0.009927817  0.29032868  0.17731895  0.2487860  0.1959857 0.21779502
z stat                      -0.0165929460   0.0182995996                  2.15556058 -0.390595718  0.27797726 -0.13655971 -0.7831718 -0.6701045 0.09447918
p value                      0.9867613520   0.9853998469                  0.03111799  0.696096091  0.78102982  0.89137883  0.4335263  0.5027912 0.92472852
                 ccnd2      cdkn2a        e2f2       e2f3       jak1      adam10      acvrl1      aurka     chek1        dab2     eif4e       foxo1     itgav
Coefficient 0.04956548 -0.29511039 0.625570631 0.07406947 -0.1798137 -0.05394274 0.001289300 0.47187037 0.2269085 -0.35954174 0.2756818 -0.02894359 0.1060196
Std. Errors 0.19348783  0.17570765 0.232171934 0.17878285  0.1930333  0.16428125 0.203092279 0.23641166 0.2327616  0.20520727 0.2023412  0.19682675 0.1697805
z stat      0.25616845 -1.67955338 2.694428302 0.41429854 -0.9315162 -0.32835605 0.006348348 1.99596912 0.9748539 -1.75209068 1.3624604 -0.14705109 0.6244512
p value     0.79782077  0.09304424 0.007050949 0.67865548  0.3515866  0.74264247 0.994934785 0.04593728 0.3296328  0.07975822 0.1730526  0.88309169 0.5323313
               pdgfrb       rheb   rps6ka2     tgfbr2    adgra2        ctcf     fancd2 hsd17b11
Coefficient 0.3851251 -0.1132305 0.27054541 -0.2971511 0.3067531 0.008167561 -0.0574864 0.3820272
Std. Errors 0.2213317  0.1931664 0.16234939  0.2987293 0.2043368 0.145718357  0.1967805 0.2525122
z stat      1.7400360 -0.5861809 1.66643932 -0.9947172 1.5012128 0.056050326 -0.2921347 1.5129057
p value     0.0818527  0.5577540 0.09562595  0.3198738 0.1333005 0.955301716  0.7701836 0.1303036
```

The first modified model with selected variables included the variables with a significant p-value. The variables included were survival_years, age_at_diagnosis, type_of_breast_surgery, nottingham_prognostic_index, e2f2, and aurka. The first confusion matrix is for the training data set and the second confusion matrix is for the test set. Like the first model, it is able to accurately classify a significant proportion of the data, but not all of it. The training set had an accuracy of 66.92% while the test set had an accuracy of 64.86%. The accuracy is slightly lower than the original model, and the difference between the training and test set is wider. However, this model is much more parsimonious.

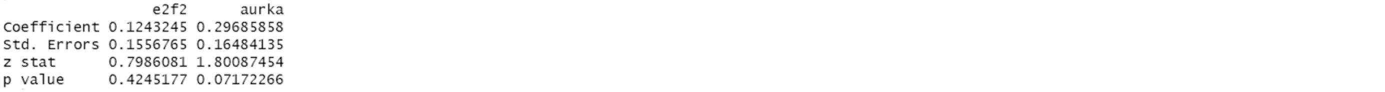


```
                     Died of Other Causes Living Died of Disease
Died of Other Causes                   79     68              43
Living                                 39    308              54
Died of Disease                        28     67             218
```



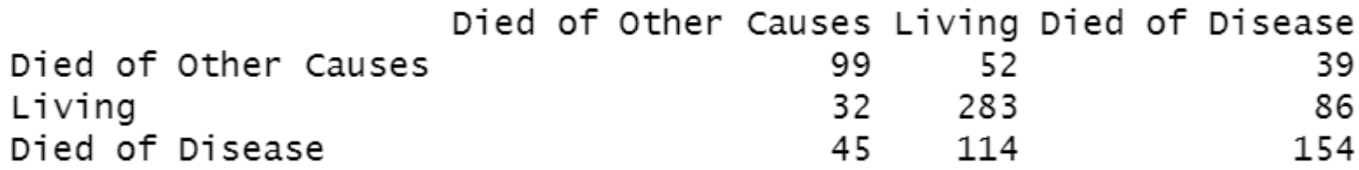
```
                     Died of Other Causes Living Died of Disease
Died of Other Causes                   52     34              23
Living                                 13    109              16
Died of Disease                        16     34              90
```

Once again, the coefficients, standard errors, z-scores, and p-values were computed and put into a table. While all of the clinical variables have significant p-values, the two gene variables do not.

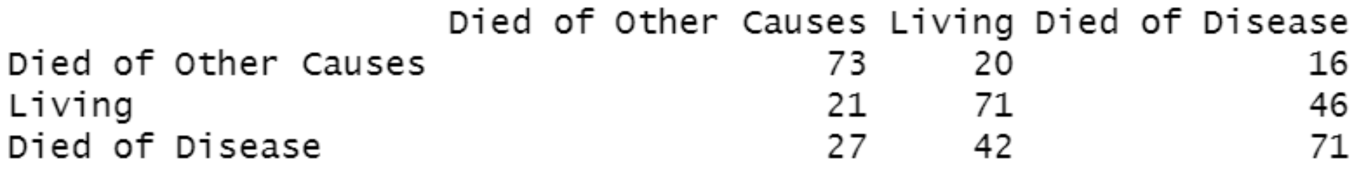


```
            (Intercept) survival_years age_at_diagnosis type_of_breast_surgeryBREAST CONSERVING type_of_breast_surgeryMASTECTOMY nottingham_prognostic_index
Coefficient 4.463865e+00    -0.17141525      -0.09728277                            2.219082e+00                     2.244784e+00                0.4075753298
Std. Errors 6.220783e-01     0.02033918       0.01077931                            3.136880e-01                     3.490100e-01                0.1166187536
z stat      7.175729e+00    -8.42783341      -9.02494997                            7.074169e+00                     6.431860e+00                3.4949381384
p value     7.192025e-13     0.00000000       0.00000000                            1.503464e-12                     1.260514e-10                0.0004741716
                 e2f2      aurka
Coefficient 0.1243245 0.29685858
Std. Errors 0.1556765 0.16484135
z stat      0.7986081 1.80087454
p value     0.4245177 0.07172266
```

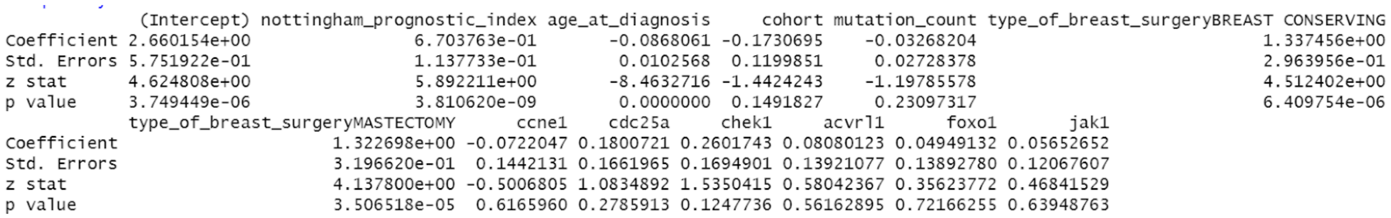
Next, Evan tried to create a model using the variables that were found to be significant in the models of the rest of the group. The clinical variables include the nottingham_prognostic_index, age_at_diagnosis, cohort, mutation_count, and type_of_breast_surgery while the gene variables are ccne1, cdc25a, chek1, acvrl1, foxo1, and jak1. This model did not perform nearly as well. With the training set, there was an accuracy of 59.29%. The model's prediction of the test set had an accuracy of 55.56%. The first confusion matrix is for the test data and the second is for the training. Based on the p-values, the variables that were most significant were the type_of_breast_surgery, nottingham_prognostic_index, and age_at_diagnosis, which were all clinical variables.



```
                     Died of Other Causes Living Died of Disease
Died of Other Causes                   99     52              39
Living                                 32    283              86
Died of Disease                        45    114             154
```



```
                     Died of Other Causes Living Died of Disease
Died of Other Causes                   73     20              16
Living                                 21     71              46
Died of Disease                        27     42              71
```



```
            (Intercept) nottingham_prognostic_index age_at_diagnosis     cohort mutation_count type_of_breast_surgeryBREAST CONSERVING
Coefficient 2.660154e+00                6.703763e-01       -0.0868061 -0.1730695    -0.03268204                           1.337456e+00
Std. Errors 5.751922e-01                1.137733e-01        0.0102568  0.1199851     0.02728378                           2.963956e-01
z stat      4.624808e+00                5.892211e+00       -8.4632716 -1.4424243    -1.19785578                           4.512402e+00
p value     3.749449e-06                3.810620e-09        0.0000000  0.1491827     0.23097317                           6.409754e-06
            type_of_breast_surgeryMASTECTOMY      ccne1    cdc25a     chek1     acvrl1      foxo1       jak1
Coefficient                     1.322698e+00 -0.0722047 0.1800721 0.2601743 0.08080123 0.04949132 0.05652652
Std. Errors                     3.196620e-01  0.1442131 0.1661965 0.1694901 0.13921077 0.13892780 0.12067607
z stat                          4.137800e+00 -0.5006805 1.0834892 1.5350415 0.58042367 0.35623772 0.46841529
p value                         3.506518e-05  0.6165960 0.2785913 0.1247736 0.56162895 0.72166255 0.63948763
```

With these initial models, we get a good idea that the "clinical" variables tend to be much better predictors of the death outcome. While certain genes such as aurka may play a role in prediction, their significance is less than other variables. Moving forward, we will be using the model that was based on the p-values of the original, which was the first modified model that we have shown. This is because it is simple to interpret while still having as high of an accuracy as the original.

One of the biggest takeaways from this analysis was that the clinical variables such as the type of surgery, age, and the Nottingham Prognostic Index play a significantly bigger role in determining the outcome of dying from cancer than the gene variables do. Moving forward with this, it may be best to research other clinical variables that play a role in surviving breast cancer. Since the models created here were about 70% accurate at the highest, there must be other variables at play that can help improve accuracy of prediction.