Final Project Report: Breast Cancer Research

Vaishnavi Shastri (vshastri@iu.edu)
Palavi Dhanaji Patil (palpatil@iu.edu)
Tumul Rajvedi (trajvedi@iu.edu)
Tejas Padmanabhan (tpadman@iu.edu)

1. Statement of goals

Questions we are trying to answer:

Does HER2 and Tumor Stage play an important role in the patient's survival or does the survival also depend on other factors?
a) How Tumor Stage with protein 4 affects the survival of the patients?
b) How Tumor Stage with protein 4 affects the survival of the patients with Age?

Importance of exploring Breast Cancer Dataset:

Detailed study of breast cancer data is essential for advancing our understanding of factors that significantly impact survival rates. By identifying how elements like HER2 status, Tumor Stage, and Protein4 levels influence outcomes, we can improve treatment protocols and offer more personalized approaches that enhance patient prognosis. T Given that breast cancer is one of the most common cancers globally, this is not only important for clinical relevance, but it also shapes evidence-based practices that enhance patient quality of life. Such analyses reveal patterns that may not be apparent from less thorough studies, contributing to a robust body of knowledge that supports more effective strategies for managing breast cancer.

Furthermore, this research has significant implications for public health. It empowers healthcare providers to better advise on personalized prevention and treatment options, thus potentially reducing healthcare costs and improving outcomes. The insights gained also aid in crafting guidelines that boost survival rates, underscoring the societal and economic importance of targeted cancer research. This commitment to in-depth analysis and application of its findings is essential for pushing the boundaries of current medical practice and improving overall patient outcomes in breast cancer treatment.

2. Data Description and Graphs

Data Set Link: https://www.kaggle.com/datasets/amandam1/breastcancerdataset

Verbal Description:

This dataset consists of a group of breast cancer patients, who had surgery to remove their tumor. Data was collected by *Queen's University Belfast Cancer Research Centre.* We have chosen the dataset from kaggle. This dataset is appropriate for a variety of data analysis and modeling applications in the healthcare domain because each column contains detailed information about patients and the type of breast cancer they had as well as other detailed information regarding the patient's tumor. The dataset in question has a total of 16 columns (variables) and 300+ records. The dataset is collected over the range of 4 years from 2017 to 2021.

Patient_ID: unique identifier id of a patient

Age: Age at diagnosis (Years)

Gender: Values will be Male/Female

Protein1, Protein2, Protein3, Protein4: protein structure levels (higher the protein number, the more structured it is)

Tumour_Stage: I, II, III (Stage at which Cancer has grown to)

Histology (Type of Cancer) : Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, Mucinous Carcinoma

ER status (Estrogen Receptors): Values can be positive or negative meaning it can either receive or not receive signals from estrogen telling it to grow.

PR status(Progesterone Receptors): Values can be positive or negative meaning it can either receive or not receive signals from progesterone telling it to grow.

HER2 status (Human Epidermal Growth Factor Receptor 2): Values can be positive or negative indicating whether the protein is growing the size of the cancer cells (positive) or not (negative).

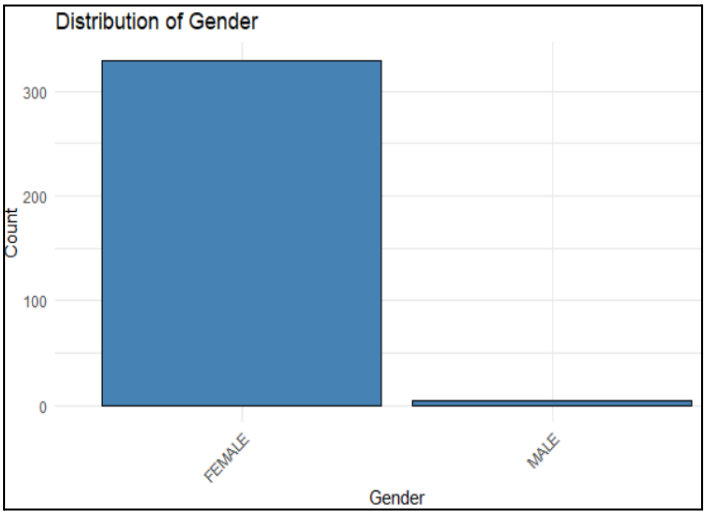
Surgery_type: Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other

Date_of_Surgery: Date on which surgery was performed (in DD-MON-YY)

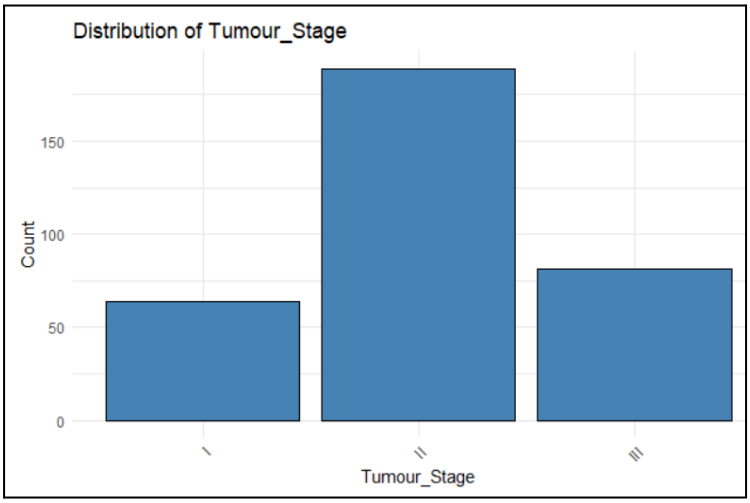
Date_of_Last_Visit: Date of last visit (in DD-MON-YY) [can be null, in case the patient didn’t visited again after the surgery]

Patient_Status: This is our response variable. Values can be Alive or Dead

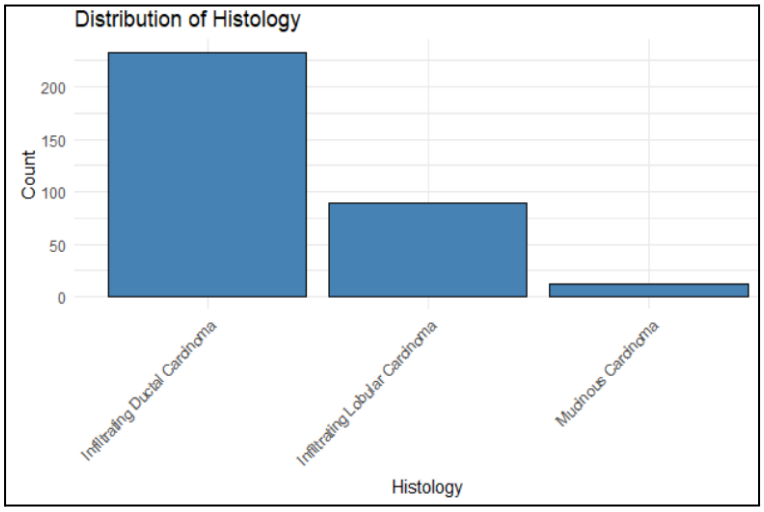
Graphical Description:



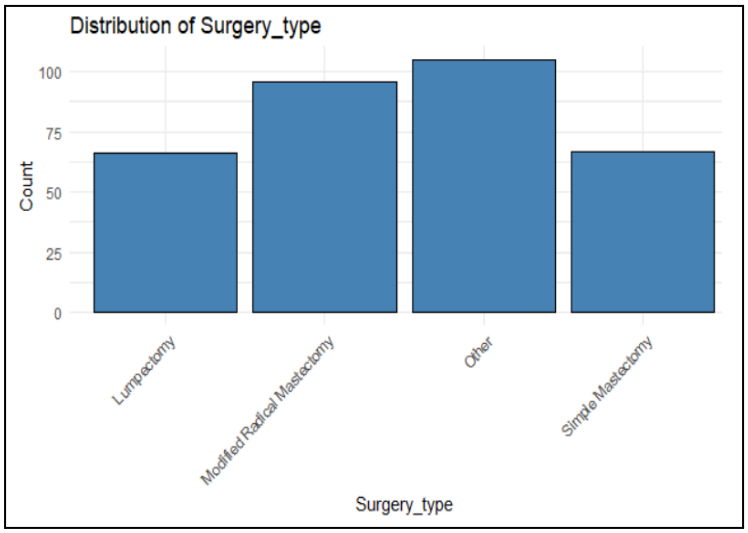
As observed in this distribution graph for Gender, the dataset has more records for female than male cancer patients. This shows women are more susceptible to breast cancer than men.



With respect to tumor stage, stage II patients are more prevalent in the dataset than other stages. For Surgery type, “Other” types of surgery is preferred by the cancer patients.



In the distribution of histology, it can be seen that Infiltrating Ductal Carcinoma is more common amongst patients followed by Infiltrating Lobular Carcinoma and then Mucinous Carcinoma.

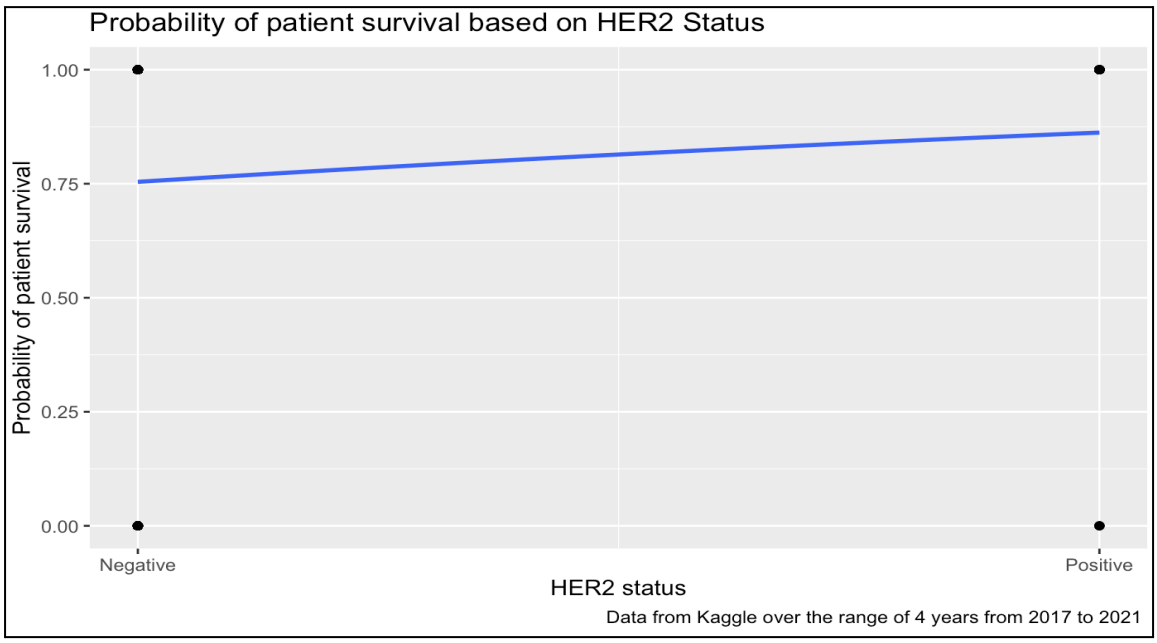


Patients may opt for "Other" types of surgeries due to their personal preferences which may include prioritizing to maintain a natural appearance with minimal scars. Additionally, patients enrolled in clinical trials may undergo experimental surgical procedures that offer innovative options for treating breast cancer, further expanding the range of surgeries classified as "Other."

3. Models and Results

In our breast cancer study, we employed GLM[1], due to its capacity to model binary outcomes like patient survival (Alive/Dead), offering a good fit to the non-normal distribution of our data.

I) Patient survival based on HER2 status



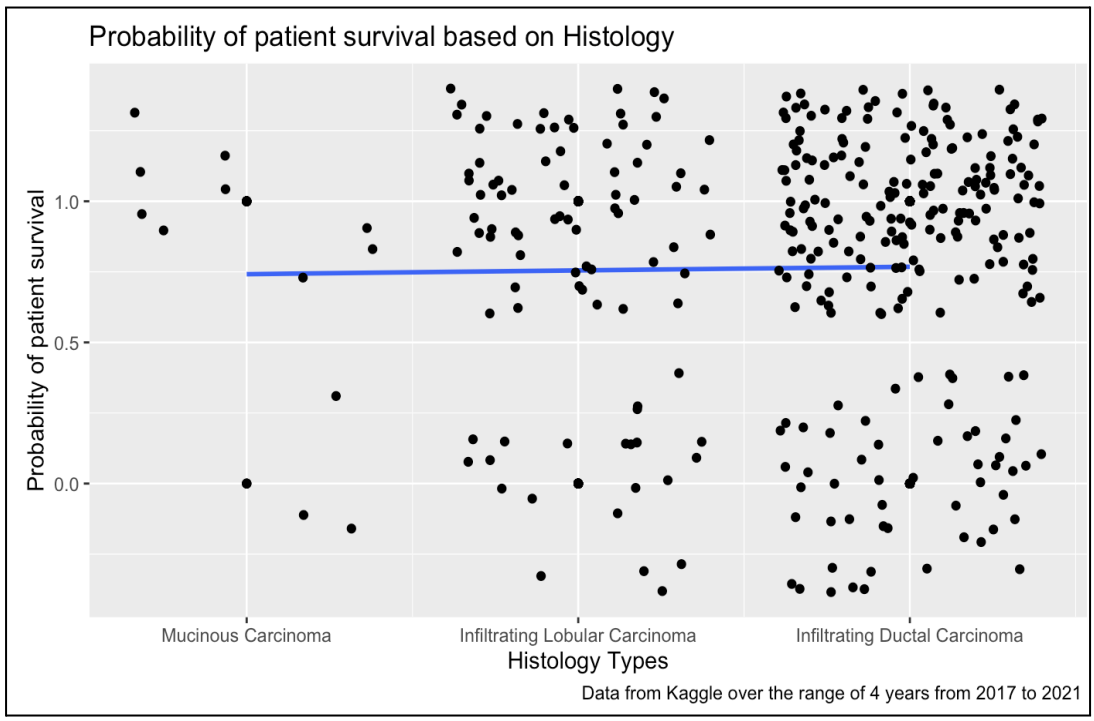
[1]

From the above graph, we observe there is a slight increase in survival probability for patients with positive HER2 status compared to negative. However, the close proximity of the data points near the top suggests that HER2 status alone may not be a strong predictor of survival in this dataset. Further exploration is shown is appendix. [3]

Hence, we decided to explore other factors that may affect the Survival probability.

II. Patient survival based on Histology

First factor we decided to explore is the effect of different histology types on the probability of patient survival.



[2]

We observed that overall there are negligible changes in survival rates, suggesting other factors besides histology influence patient survival outcomes. However, infiltrating ductal carcinoma cancer type has slightly more survival rate than infiltrating lobular carcinoma and mucinous carcinoma, but this change is not significant enough to tell more about survival outcome of patients.

III. Patient survival based on Tumor stages:

Our next choice was to investigate the average survival rates across various tumor stages to ascertain the influence of tumor stage progression on patient survival, as it is well known that tumor stages have an impact on patient's survival.

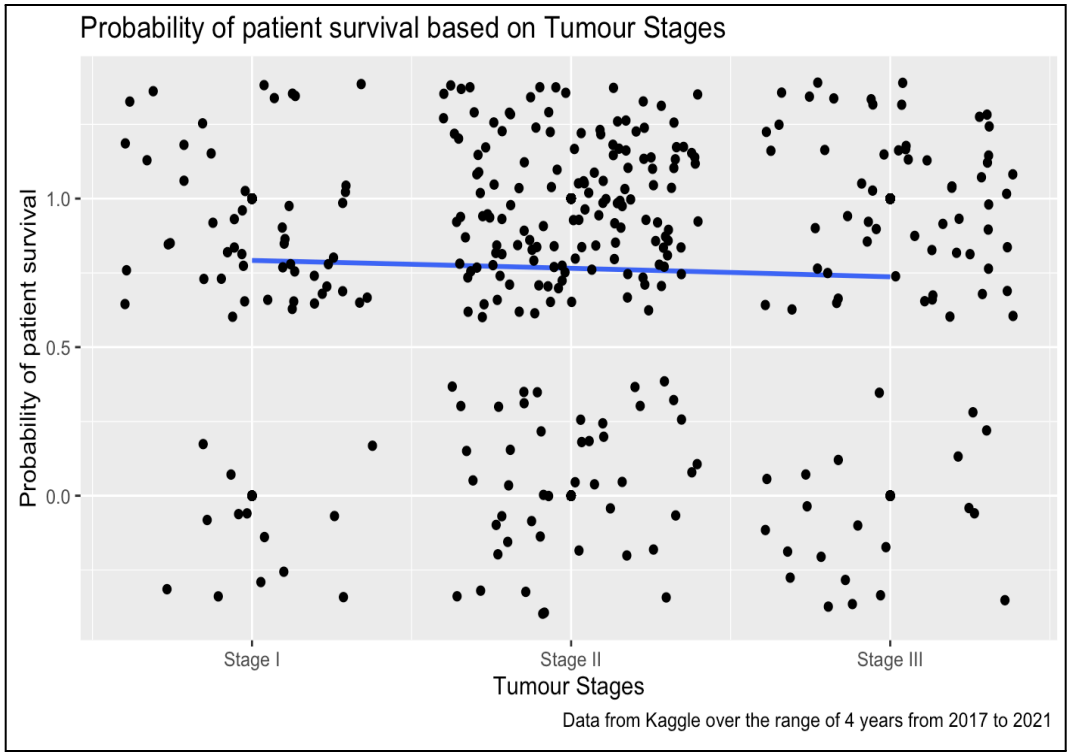
a) The Average survival rate for every stage:

Stage	Survival Chances in Percentage
Stage I	79%
Stage II	76%
Stage III	74%

The calculated average, shows a marginal decline in survival rates with advancing tumor stages. However, the relatively small range of this decline suggests that while tumor stage is a factor in patient survival, it is not the sole determinant and other factors contribute to a patient's survival. The following graph provides further illustration of this trend.

b) Graphical representation:

[3]



These findings directed our analysis to investigate other contributing factors in conjunction with tumor stage for a comprehensive understanding of patient survival in breast cancer.

IV. Patient Survival based on Protein values

Proteins are fundamental to the development, progression, and treatment of breast cancer due to their involvement in critical cellular processes. Gaining insights into how protein structure influences breast cancer can pave the way for the development of medical strategies targeted at specific molecular changes.

Hence, our next analysis focused on calculating the Bayesian Information Criterion (BIC^2) values for interactions between various proteins and tumor stages. This was done to identify the most statistically significant predictors among the proteins when considering their interaction with tumor stages.

a) BIC^2 values calculation of all the proteins with Tumor stage interaction:

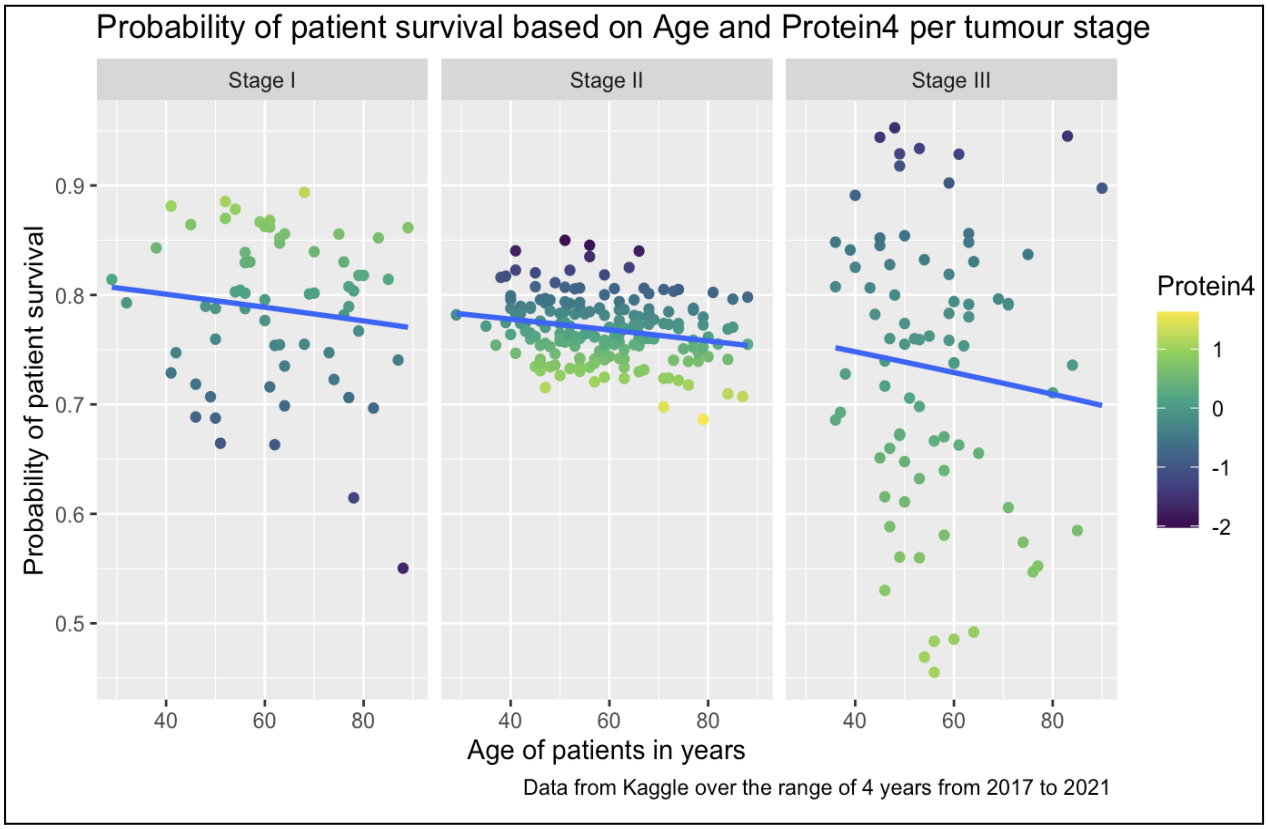
Protein and Tumor stage interaction	**BIC values**
Protein1 + Tumour_Stage	382.0652
Protein1 * Tumour_Stage	386.9556
Protein2 + Tumour_Stage	380.9194
Protein2 * Tumour_Stage	386.6094
Protein3 + Tumour_Stage	381.8195
Protein3 * Tumour_Stage	386.9689
Protein4 + Tumour_Stage	380.2533
Protein4 * Tumour_Stage	378.5631

From the BIC^2 values, the interaction between Protein 4 and Tumor Stage presents the lowest BIC^2 score (378.5631), indicating that this combination provides the most efficient model as lower BIC^2 values imply that the model is neither overfit nor excessively complex relative to its predictive ability.

While the BIC^2 values for interactions involving Protein2, Protein3, and even the additive effect of Protein1 and Tumor Stage are relatively low, they are consistently higher than those observed for Protein 4 interactions.

This pattern highlights Protein4's stronger predictive power regarding survival probabilities when analyzed in conjunction with Tumor Stage. Hence we decided to use the interaction of protein 4 and tumor stage for our predictive models.

b) Adding Age as a dependent factor with Protein4 and Tumor stage to predict the probability of patients' survival and plotting the graph for the same.



[4]

Observations:

Age V/s Survival: Survival rates decrease as patients' age increases across all tumor stages.

Stage I Survival: Higher survival rates are associated with higher levels of Protein 4, indicated by light green points in Stage I.

Early-stage Tumor Response: The less aggressive nature of Stage I tumors may be more responsive to treatment, potentially linked to the structural benefits of Protein 4.

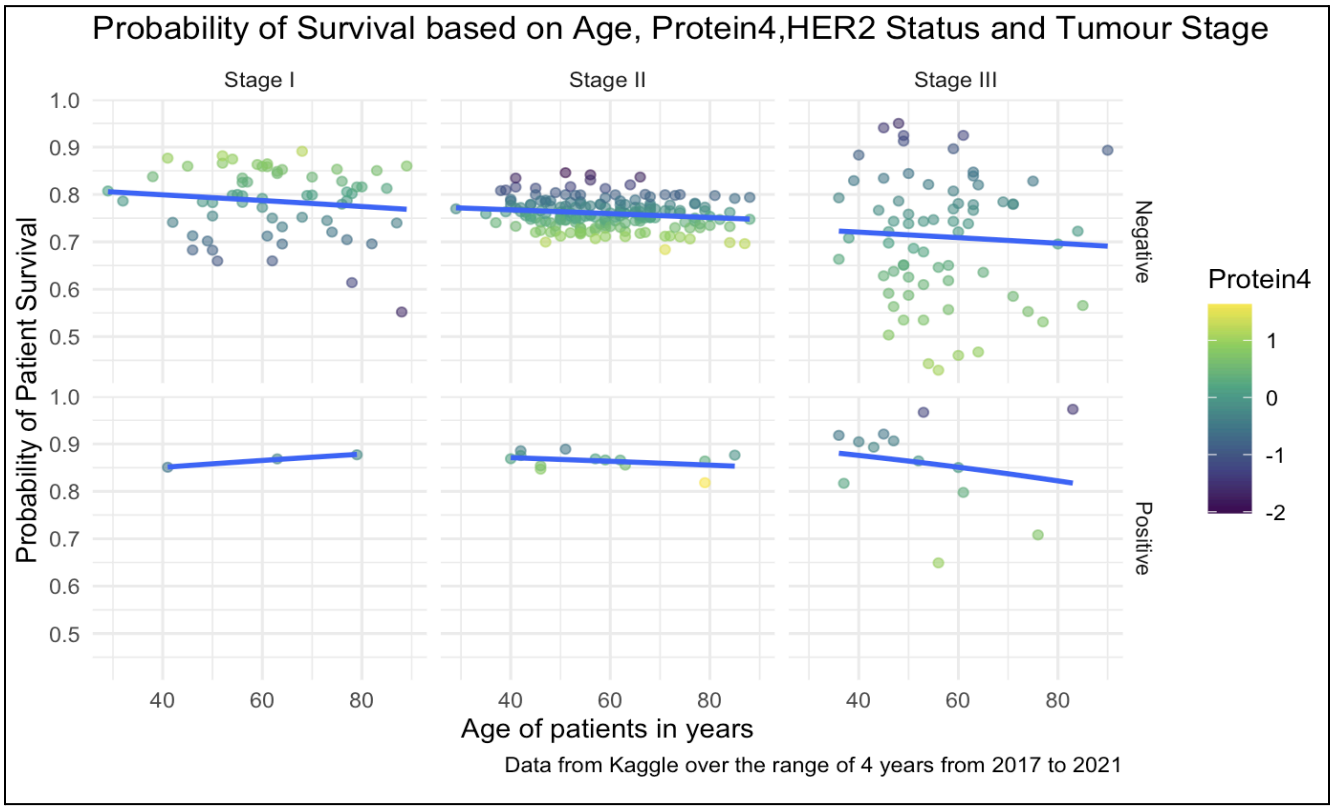
Tumor Progression: As tumors progress to Stages II and III, the survival rates drop regardless of Protein 4 levels.

Advanced Stages: In later tumor stages, the aggressiveness of cancer diminishes the influence of Protein 4 structure on survival rates.

Influence on Prognosis: Both age and tumor stage heavily impact survival probabilities, with protein structure having a potential prognostic role in early-stage breast cancer.

Protein 4's Role: Protein 4 may contribute to better outcomes in initial stages, but its impact lessens as tumor stages become more aggressive.

c) We have plotted the graph for a multifaceted analysis of survival probabilities in breast cancer patients, integrating the effects of HER2 status, age, Protein4 levels, and tumor stages.



[5]

Observations:

HER2 Status: Unexpectedly, the graph shows that patients with positive HER2 status have higher survival probabilities across all stages, contrary to typical expectations. This anomaly may be attributed to a lack of sufficient data for positive HER2 status, which could skew the results.

Age: The data suggests a trend where survival probabilities diminish with increasing age in all three tumor stages, with the exception of Stage I for positive HER2 status. The exception is likely due to an insufficient number of data points for patients with positive HER2 status in Stage I, which could distort the true pattern.

Protein4 Levels: In Stage I, there's an observed increase in survival probability with higher Protein4 levels, which can be linked to easier detection and treatment at this early stage. However, this pattern does not hold in Stages II and III, where increased Protein4 levels do not correspond to better survival chances, suggesting that the complexity of treatment increases with the tumor stage.

Tumor Stage: For patients with negative HER2 status, there is a noticeable decrease in survival probability as the tumor stage progresses from I to III, indicating that higher-stage cancers are more challenging to treat and have worse survival rates.

Overall, these observations suggest that while certain patterns emerge, the interaction between these variables is complex.
The presence of higher Protein4 levels seems beneficial in early-stage tumors but less so in later stages. Age consistently appears to influence survival negatively. The decrease in survival with more advanced tumor stages highlights the need for aggressive and early treatment strategies.
However, the lack of data for positive HER2 status across all stages limits our analysis.
In conclusion the model points to the multifactorial nature of breast cancer survival, with tumor stage, age, and protein structure all playing critical roles.

4. Conclusions, limitations, & future work

Conclusion:

1. HER2 status, Tumor stage and Histologies alone are not enough to predict the survival status of the patient.
2. The predictions cannot be calculated based on HER2 status only, as the data points for the HER2 column values are unbalanced (i.e Negative HER2 has more values as compared to Positive HER2)
3. The Rate of Survival varies very little in all the three stages hence (decreases only by 2% as stage increases) hence only tumor stage cannot be used to predict the Survival.
4. Increase in protein4 value increases the chances of survival in stage I while it decreases the chances of survival for stage II and stage III.
5. In all the three stages , the age as the parameter plays the crucial role to predict the survival probability of the patient. As the age increases, survival chances decrease for all the three stages.
6. In conclusion, Patient survival not only depends on HER 2 and tumor stage, but parameters like protein4 and age of patient also have a significant role in deciding the chances of survival.

Limitations:

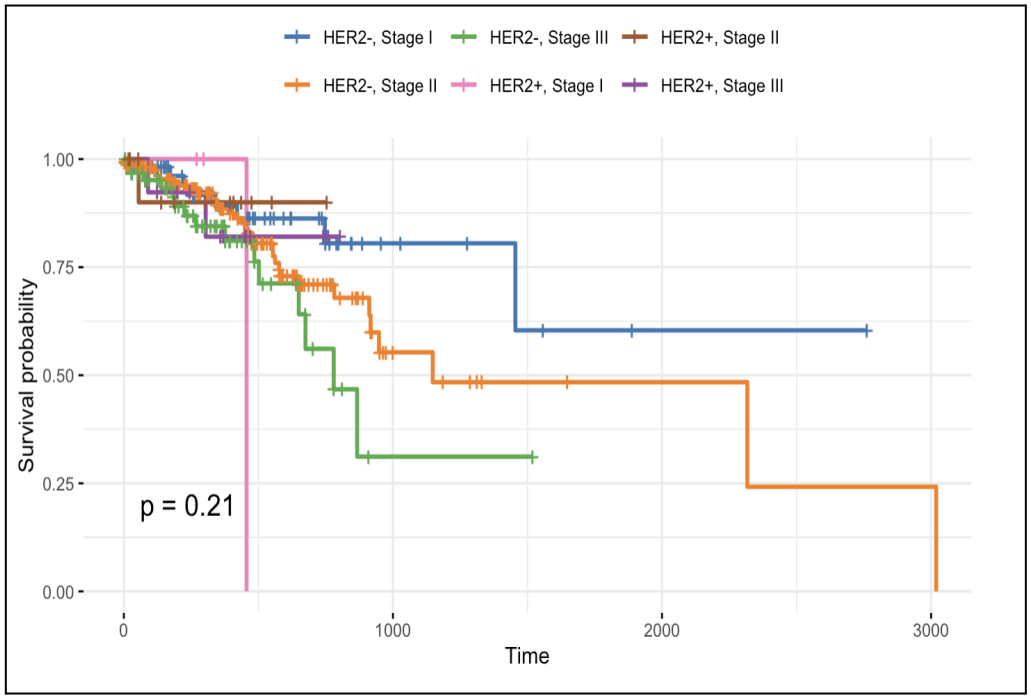
1. A limitation encountered during the exploratory data analysis (EDA) for this project was the limited size of the data set, particularly the insufficient number of records for male patients with breast cancer. This constraint limited the ability to conduct a meaningful EDA with respect to gender.
2. Another limitation was the imbalance in HER2 status within the dataset. The higher occurrence of "Positive" values compared to "Negative" ones resulted in an unbalanced dataset, potentially impacting the analysis and any subsequent conclusions.
3. Lastly, some records had blank values for the response variable (Patient Status). To maintain the integrity of the analysis, these records were removed from the dataset.

Future Work:

1. For future analysis, further investigation into HER2 status could be beneficial. Graph 5 suggests that the probability of patient status increases for "Positive" HER2 status, a trend that diverges from other GLM[1] lines in the model. The underlying reasons for this trend could be explored in future studies.
2. Another area for future exploration could involve examining the impact of surgery types. The dataset indicates a preference for the "Other" category of surgeries among patients. Further research could provide insights into the characteristics of this category and its influence on patient survival outcomes.

5. Appendices:

1. The Regression Model Used for prediction is Logistic Regression 'GLM'. It is a statistical method used for modeling the relationship between one or more independent variables (predictors) and a categorical dependent variable (outcome) with two possible outcomes, typically coded as 0 or 1. It's commonly used for binary classification tasks.
2. BIC (Bayesian Information Criterion) score is used to select the best-fitting model from a set of parameters and provides a balance between model goodness of fit and simplicity. Lower BIC scores indicate better model fit, and the model with the lowest BIC score is typically chosen as the preferred model.
3. Kaplan-Meier survival plot:



The graph depicts the probability of survival (y-axis) at different times after diagnosis or the start of treatment (x-axis).Each line represents a combination of HER2 status and Tumor Stage.
Censoring (+) on the curves indicate 'censored' data, which occur when a patient leaves the study before an event (death) occurs or the study ends before the patient has an event.
The overlapping survival curves and non-significant p-value (p = 0.21) indicate that HER2 status does not appear to have a substantial impact on survival in this dataset, even when combined with Tumor Stage.

4. Patient survival based on Protein4, HER2 and Tumor stages: