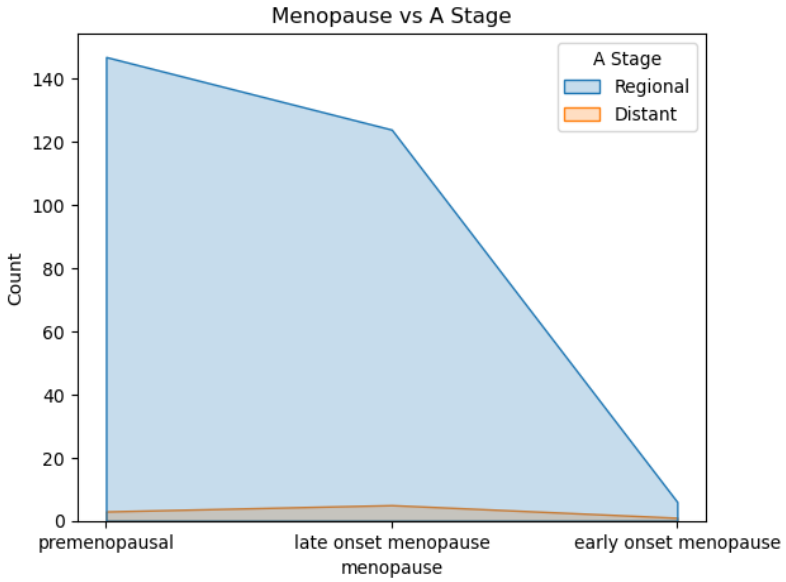
Miranda Smith
Molleigh Hughes
Tasha Christensen
Hidy Vengalil
Stephanie Santiago

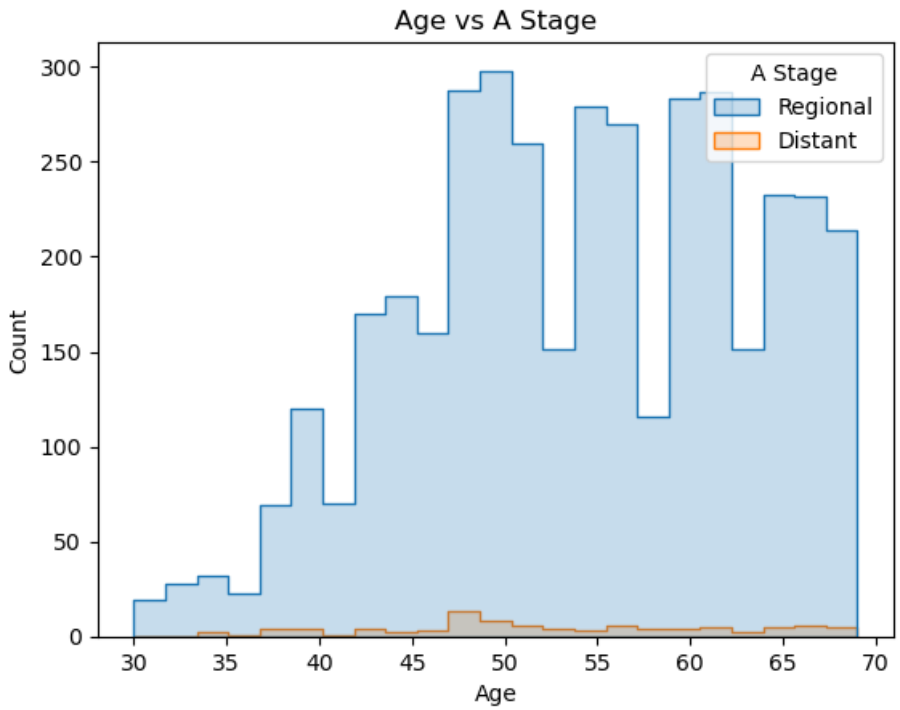
Survival Analytics: Charting the Course of Breast Cancer Progression

Our project was motivated by an interest in women's health issues; with that in mind, we decided to focus our analytical lens on breast cancer. That led us to examine how different stages of a woman's life and different avenues of disease impacted the severity of breast cancer and impact on survival rates. With the following analysis, we sought to understand relationships between metastasis, menopausal status, age, survival length, and recurrence.

Menopause & Age and Axillary Nodes (A Stage)



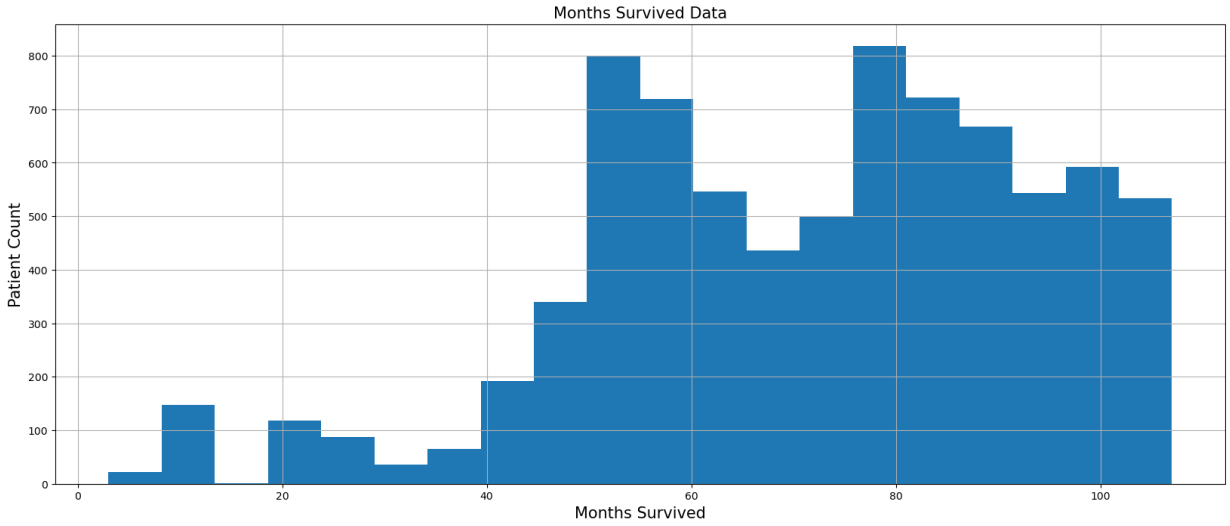
In our data, only 2% of patients exhibit cancerous axillary nodes (A Stage) in distant parts of the body. The histogram illustrates that late-onset menopausal patients are more likely to have distant cancerous nodes compared to premenopausal and early-onset menopausal women. Notably, the group of women with early onset menopause has the fewest cases of breast cancer, with just one patient showing cancer spread to remote nodes.



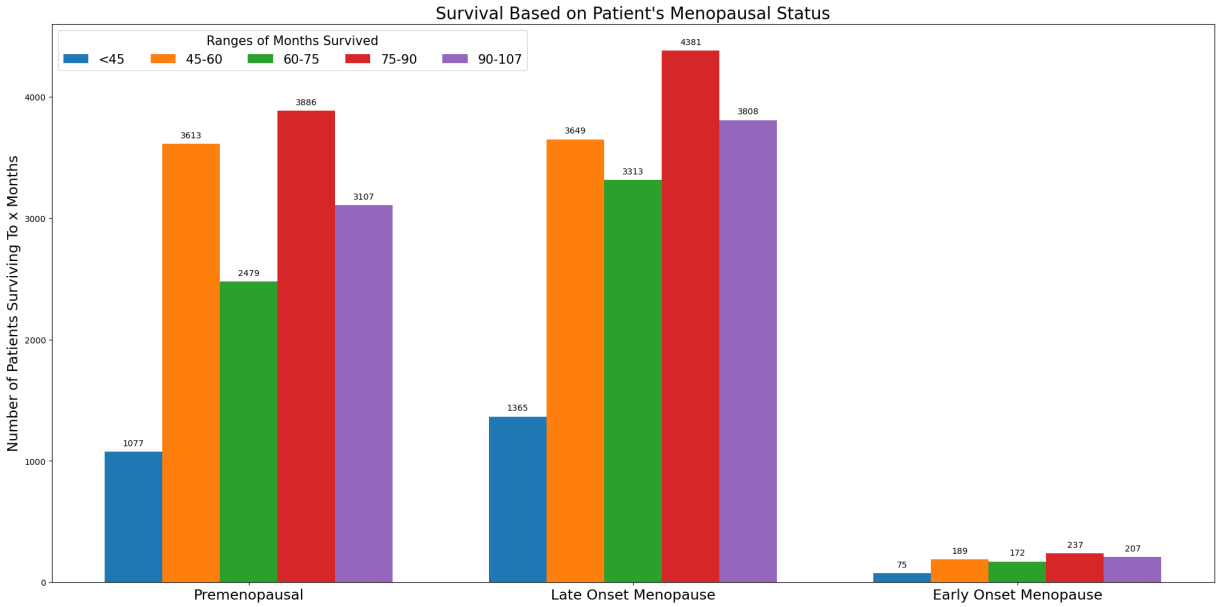
The relationship between age and metastasis in breast cancer patients reveals that those with distant axillary nodes spread encompass a wide age range from 33 to 69, with a peak occurrence at around 47 years, likely attributed to routine mammogram screening.

Menopause and Survival Months

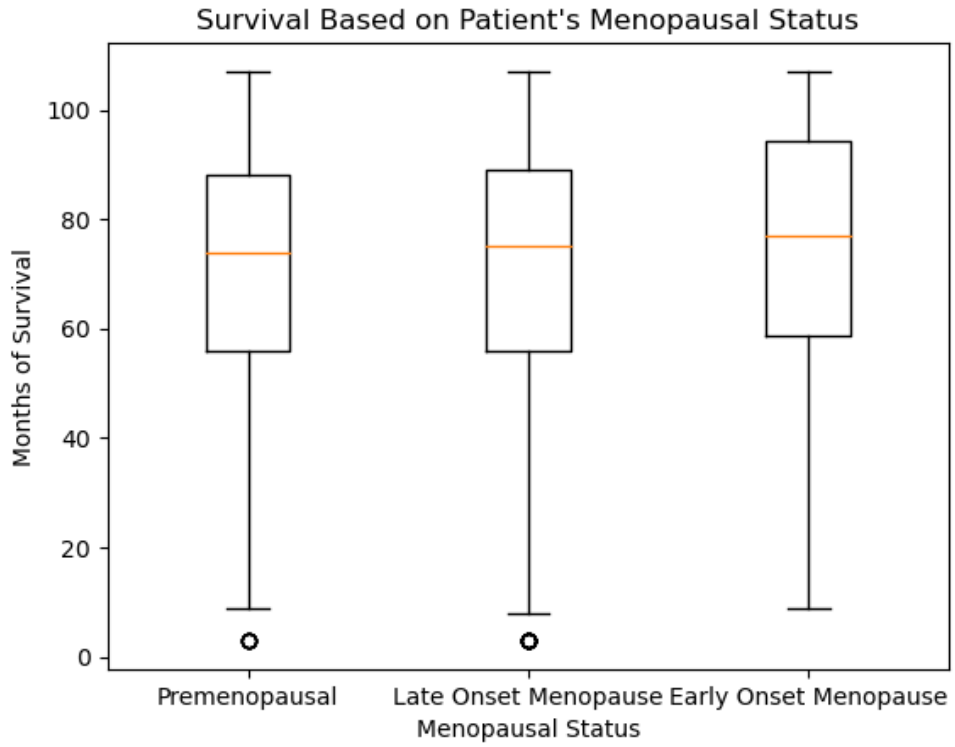
We next decided to examine the relationship between menopausal status and the survival months of the patients.



The histogram reveals a left-skewed distribution of survival months in the dataset, suggesting a non-normal distribution. This is why, when bucketing for survival months data, the first bucket includes everything under 45 months of survival.



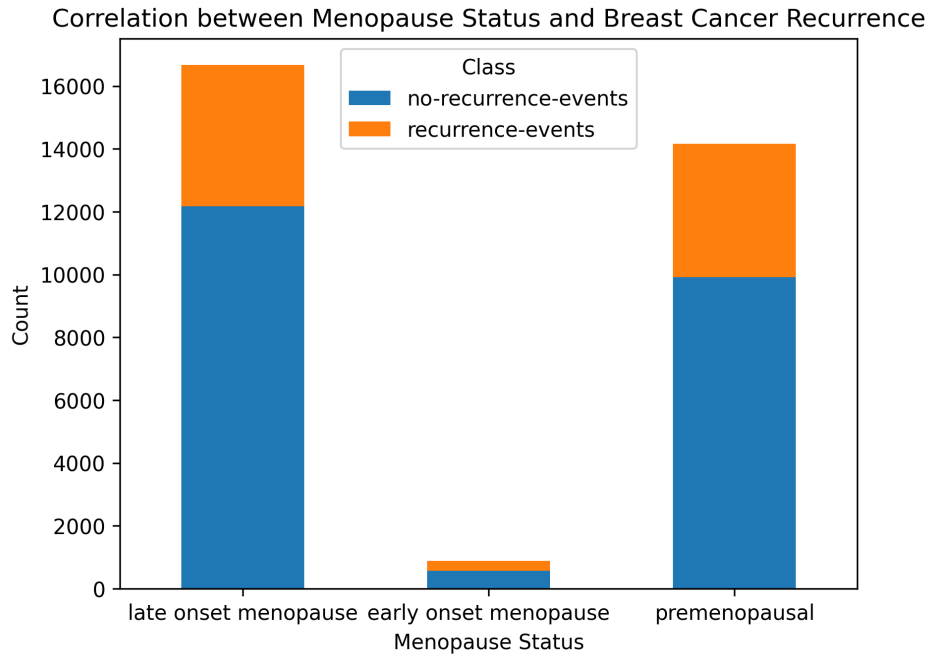
These bar charts depict the central tendencies of both survival months and patient age, with the dataset ranging from 2 to 107 months of survival and a mean of around 72, in alignment with the left-skewed nature seen in the histogram. The bar chart demonstrates that regardless of menopausal status, most breast cancer patients survive between 75 to 90 months, with very few living less than 45 months. Each menopausal status has a decrease in patients from 45-60 months to 60-75 months and then a subsequent increase in patients in the 75-90 column. Further investigation would be needed to determine the cause of this dip, whether it's only true of this data set or across all patients, but its occurrence across all three groups is notable.



Lastly, the boxplot shows that survival months among different menopausal status groups have roughly equal means, indicating similar longevity among patients regardless of menopausal status.

Menopause and Recurrence

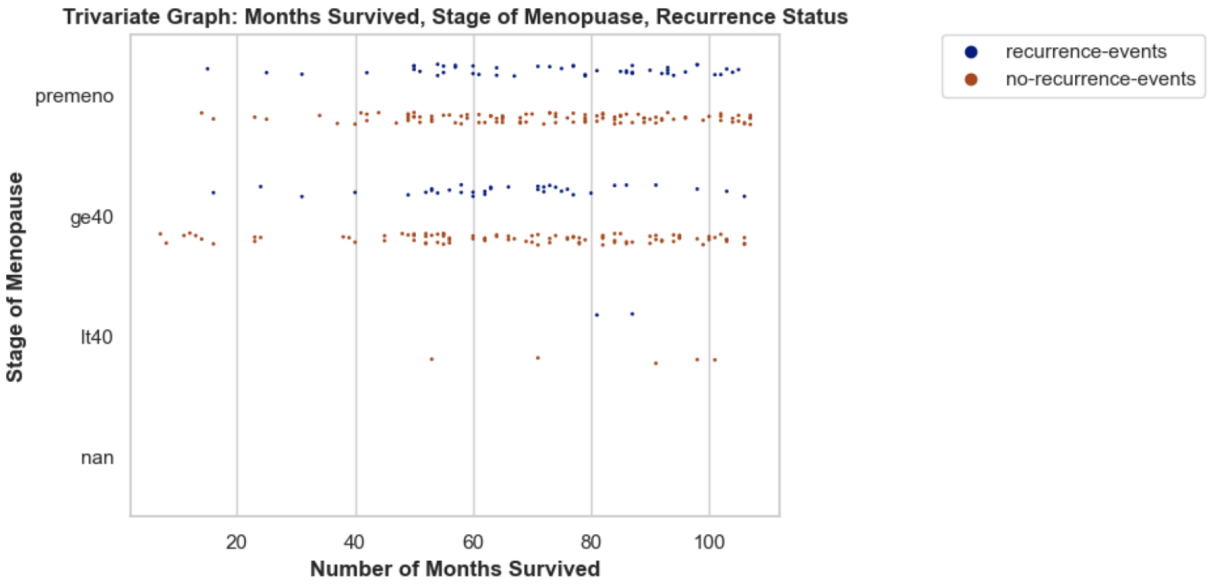
Next, we examined menopause status with the recurrence of breast cancer in the patient population. The following graphic visually represents the relationship between menopause status and recurrence using distinct bars for each menopause status, with segments indicating the counts of recurrence within each status.



A larger segment in one menopause status suggests a higher prevalence of that recurrence, while a smaller segment implies it's less common. A Chi-square test was performed, revealing a significant association between the two variables, as indicated by a large Chi-square statistic (52.87) and a very low p-value (3.31e-12). The results confirm a significant correlation between menopause' status and recurrence, suggesting that menopause status may influence the recurrence rate.

Menopause, Survival Months, and Recurrence

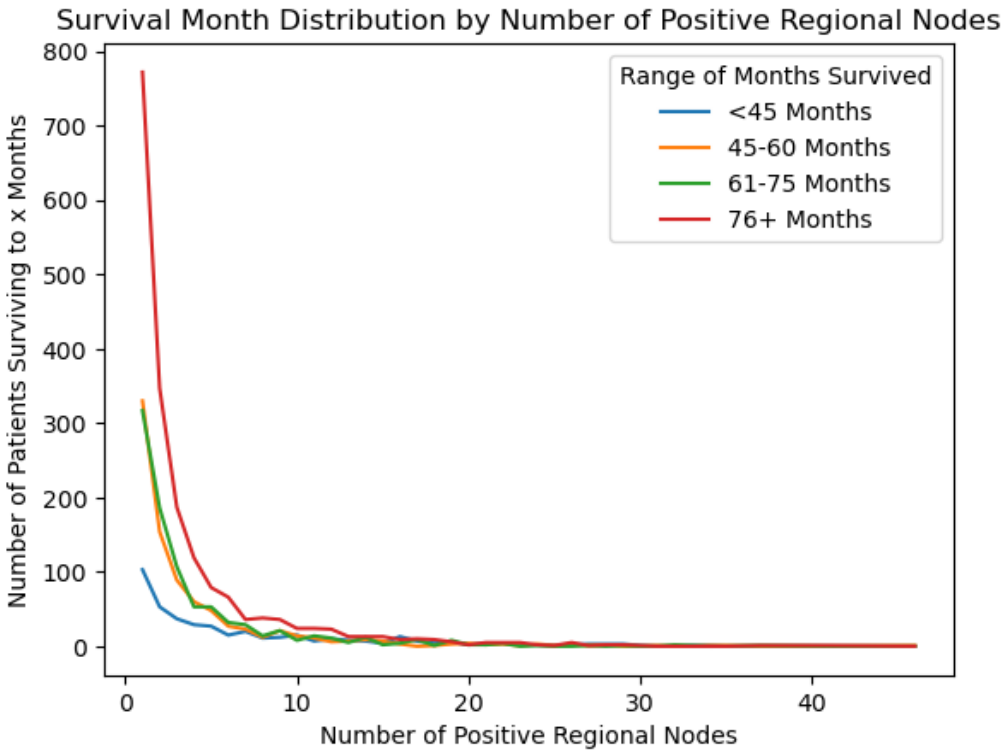
In an effort to try and gain a broader view of all of our data in conversation with itself, we created a trivariate graph, charting the relationship between menopause, months survived, and recurrence events.



Though it's difficult to point to any correlation, there are some interesting trends. The data is most concentrated for women surviving 50 months or more with no-recurrence. This might be due to the surveyed group, but it's encouraging to see the highest density of patients living for over 4 years, in comparison to other outcomes.

Positive Regional Nodes and and Survival Months

Regional lymph nodes are the nodes along the major arteries that supply blood to the cancerous part of the body; nodes testing positive for cancerous cells indicate that those cells have more access to the blood stream. With this thought in mind, knowing that metastatic occurrences can materialize through this pathway, we wanted to examine the relationship between number of positive regional node sites in a patient and survival months.



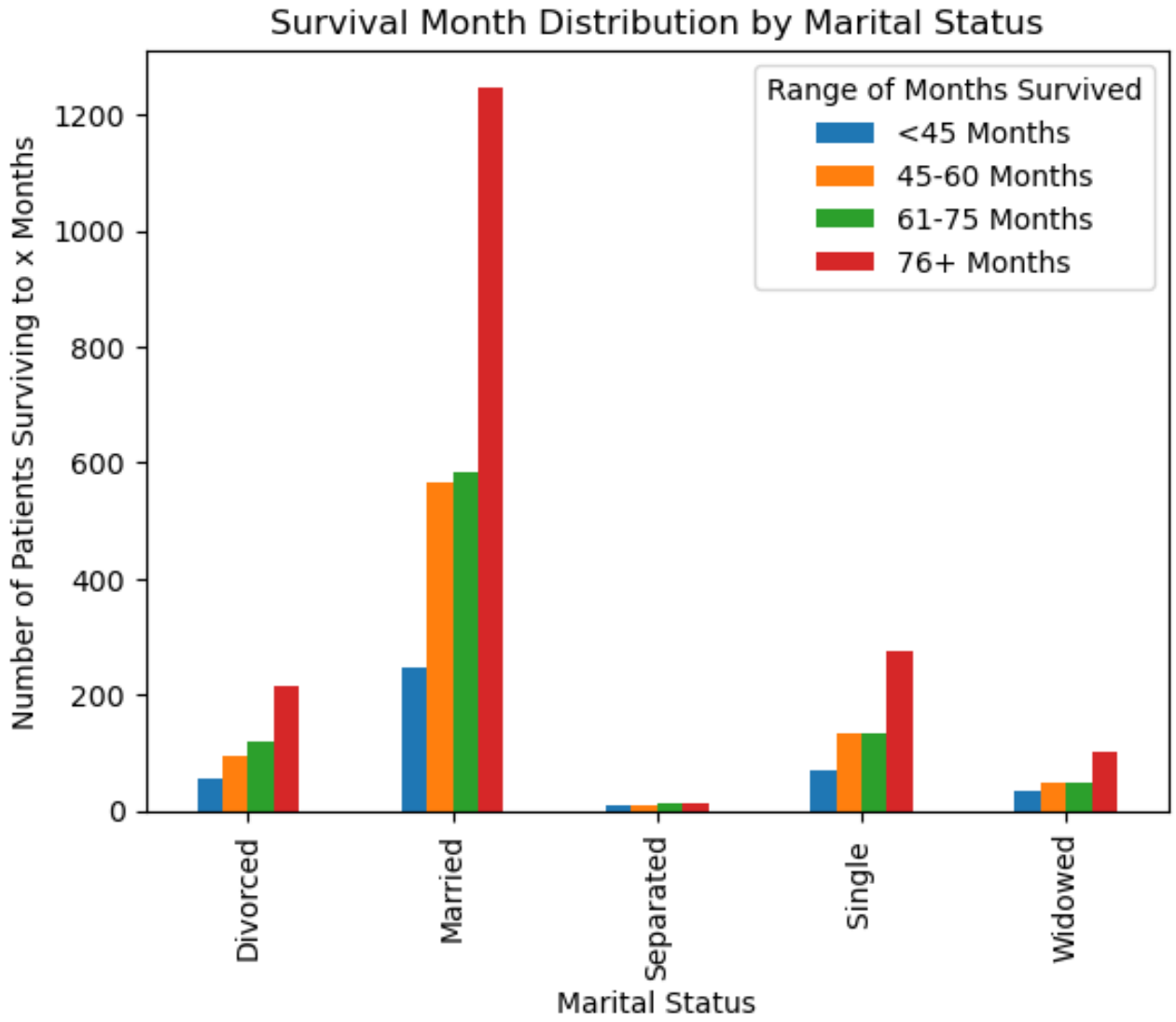
Survival Month Group	<45 Months	45-60 Months	61-75 Months	76+ Months
count	38.000000	38.000000	38.000000	38.000000
mean	10.973684	22.342105	23.710526	48.868421
std	19.117682	59.257695	60.188850	136.759242
min	0.000000	0.000000	0.000000	0.000000
25%	1.250000	1.000000	0.250000	1.000000
50%	4.000000	3.000000	2.000000	5.500000
75%	11.750000	11.750000	13.500000	24.000000
max	103.000000	330.000000	317.000000	772.000000

As illustrated in the line chart, the distribution of patients skews towards low numbers of positive regional nodes. However, the distribution under each line indicates that the fewer positive nodes a patient has identified, the more likely they survive longer than 76 months. There is a significant association between the two variables, as indicated by a very low p-value (1.56e-44).

Future Pathways of Analysis

This project was very revealing regarding how the biological changes women undergo over their lifetime impact the severity and behavior of breast cancer. In the future, it would be interesting to analyze breast cancer outcomes through the lens of social determinants of health. Things like class, race, and access to health care can impact health outcomes, especially when it comes to terminal diseases like breast cancer. In an effort to start that conversation, we analyzed survival

outcomes against marital status, hypothesizing that having a partner could potentially impact longevity due to a personal level of care in the home.



Survival Month Group	<45 Months	45-60 Months	61-75 Months	76+ Months
count	5.000000	5.000000	5.000000	5.000000
mean	83.400000	169.800000	180.200000	371.400000
std	94.336101	225.676095	230.780632	500.449598
min	10.000000	9.000000	12.000000	14.000000
25%	33.000000	50.000000	50.000000	102.000000
50%	56.000000	93.000000	121.000000	216.000000
75%	71.000000	132.000000	135.000000	277.000000
max	247.000000	565.000000	583.000000	1248.000000

As can be seen on the chart, married women make up the majority of the studied population. Due to that, it might not be fair to assume that there is association between the two variables, because while married women by far have survival ranges 76 months or over, those are also the

highest bars for women with other relationship statuses. However, a statistical test to determine the p-value indicates that these two variables most likely influence each other ($p=0.018$). This is just one example of the direction breast cancer analysis can head.