Survival Analystics: Charting the Course of Breast Cancer Progression

Data Discoveries & Breast Cancer Analysis

Overview: Project & Goals

01
Data Collection
Overview of data collection, cleanup, and exploration processes

02
Our Approach
Relevant code for analysis, unanticipated insights, problems & how resolved?

03
Results & Conclusions
Relevant images & examples that support our work. Discuss issues & attempts to resolve

04
Potential Next Steps
Potential next steps with this dataset? Data manipulations we have made? Conclusions to explore?

Exploratory Data Analysis in Healthcare

Our **exploratory data analysis** of **breast cancer research** is important for understanding the following healthcare considerations:

- Reducing time to diagnosis
- Increasing our understanding of disease risks and causes
- Developing stronger prevention strategies
- Predicting and diagnosing illness

Citing Datasets

Surveillance, Epidemiology, and End Results; (SEER)., National Cancer Institute. (2019, January 18). SEER Breast Cancer Data. Retrieved October 2023,.

Zwitter , M., & Soklic, M. (2014, April 6). Breast Cancer Data. Retrieved October 2023,.

kaggle

DATA HUB

Project & Goals

Project

For women with breast cancer, how does her **menopausal status** correlate to her **recurrence** status, **number of months survived**, and **metastasis sites**

Goals

By studying these relationships, we hope to better equip healthcare providers; for example, by **increasing understanding** of breast cancer **risks & causes**, or by **speeding up** breast **cancer predictions & diagnosis.**

Glossary Term	Description
Menopause	Cessation of menstruation and ↓ production of estrogen & progesterone; average age of natural menopause 50
lt40	**(Less Than 40)** women in reproductive years; below average age of menopause
ge40	**(Greater Than or Equal to 40)** women 40+ years old; at or past average age of natural menopause
premeno	**(Premenopausal)** women in reproductive years & not begun menopause
Metastasis sites	**(Regional v.s. Distant)** Regional site: neoplasm has not left the breast;Distant sites: neoplasm has spread beyond the breast
Survival Months	**(Integer)** Number of months the patient has survived breast cancer
Positive Regional Nodes	Metastasis sites in breast with **positive** regional nodes are most likely to form breast cancer; the fewer *positives* nodes = higher number of months survived
Recurrence	**(Recurrence v.s. No-Recurrence)** breast cancer reoccurring vs breast cancer that does not return.
Neoplasm	Abnormal and uncontrolled growth of cells within the breast tissue; typically, malignant, i.e. cancerous

01

Data Collection

An overview of the data collection, cleanup, and exploration processes

Breast Cancer Data Collection

We chose **two datasets**, because differing variables & measuring between datasets narrowed our options.

- **Kaggle** : SEER Breast Cancer Data
  - 4024 Patients in dataset
    - Survival Months (integer)
    - Metastasis Sites (Regional v.s. Distant)
- **DataHub**: Breast Cancer Data
  - 286 Patients in dataset
    - Menopausal Status (premeno, ge40, lt40)
    - Class (Recurrence v.s. No-Recurrence)

We chose these **variables**, because we wanted a sufficient amount of data, mimic complexity of future projects, and to challenge our abilities

Data Exploration

Topic

Exploration

Most interested in **women's issues**, specifically within **healthcare**

Datasets

Breast Cancer

Selecting viable data; merging on common columns

Focus

Postulate

Menopausal status & it's relationships with our chosen variables

Goal

Project

Discover meaningful correlations that empower healthcare providers

Our Cleanup Process

Step 1

Find **common column** between datasets (age)

Step 2

df_1 = (age range)
vs
df_2 = (age)

Step 3

Stephanie **converted** (age range) into (age) using mean & pushed both csv

Step 4

Molleigh called **needed columns** from each dataset & **merged** using **pd.concat**

02

Our Approach

Problems & how resolved? Relevant code for analysis?
Unanticipated insights?

Problems & Solutions

01 Data Overwhelm

Numerous women's issues in healthcare; amount of data available is overwhelming

02 Difficulty Merging

- 3 attempts to merge!
- Issues with 2 columns from each dataset,
- Numerous duplicates & NaN values skewing
- Different number of patients: 4024 patients vs 286 patients

03 No Common Column

No common columns amidst available breast cancer datasets. Thus, we converted the **age range** in one dataset to *create* our commonality: age

Main Problem: Merging

Relevant code for analysis

- 2 rows from each dataset,
- without duplicates,
- while maintaining correct total count for each variable

pd.merge, drop_duplicates(), and drop_na would not work with this data → created merge via pd.concat

	menopause	Class
count	286	286
unique	3	2
top	premeno	no-recurrence-events
freq	150	201

	Survival Months
count	4024.000000
mean	71.297962
std	22.921430
min	1.000000
25%	56.000000
50%	73.000000
75%	90.000000
max	107.000000

Example: illustrates how merged columns & kept the correct total counts for each selected variable

```
#Combine data into a single DataFrame

#Select the non-contiguous columns I want to merge from
clean_columns = clean_df[['menopause', 'Class']]

#Select the column I want to merge from unclean_df
unclean_column = unclean_df[['Survival Months']]

#Concatenate the selected columns horizontally (along columns)
trivariate_df = pd.concat([clean_columns, unclean_column], axis=

#Display theh data tabel for preview
print(trivariate_df)
```

[4]

```
...      menopause                Class  Survival Months
0       premeno     recurrence-events               60
1          ge40  no-recurrence-events               62
2          ge40     recurrence-events               75
3       premeno  no-recurrence-events               84
4       premeno     recurrence-events               50
...         ...                   ...              ...
4019       NaN                   NaN               49
4020       NaN                   NaN               69
4021       NaN                   NaN               69
4022       NaN                   NaN               72
4023       NaN                   NaN              100

[4024 rows x 3 columns]
```

```
heck our integer column
```

Unanticipated Insight

Through this research, we have learned that the recognition of breast cancer's heterogeneity has had a profound impact on the field, **emphasizing the need for personalized approaches** and **paving the way for more effective and targeted treatments.**

Link with Penn Medical research = $10 millions

grant for breast cancer recurrence

03

Results & Conclusions

Breast cancer data illustrations & examples
Issues for our illustrations & attempts resolved?

Stephanie

Menopause & Recurrence

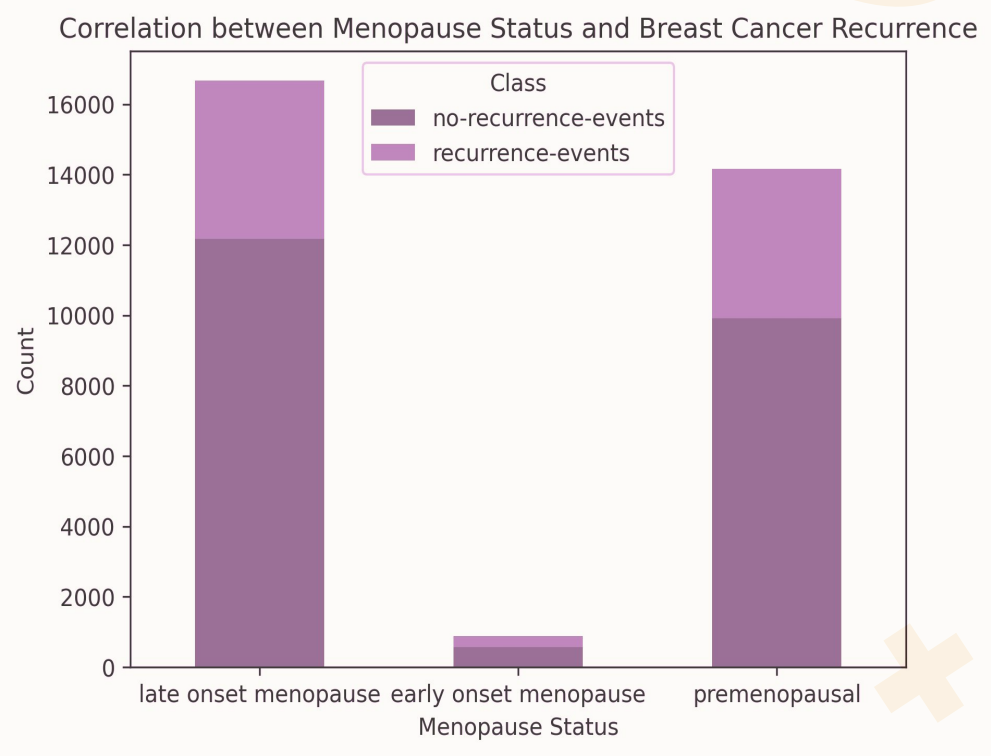
Results

P-value of 3.3084e-12 suggesting a **significant association** between **menopausal status** & **recurrence**

Conclusion

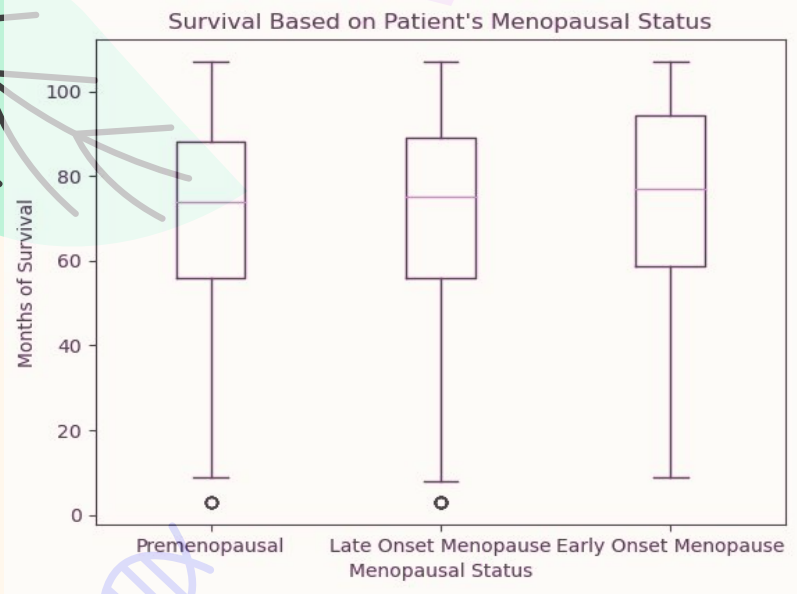
Visuals of this bar chart, suggest **~30% chance of recurrence**

Despite different sizes, the **bars** are **similarly proportioned**



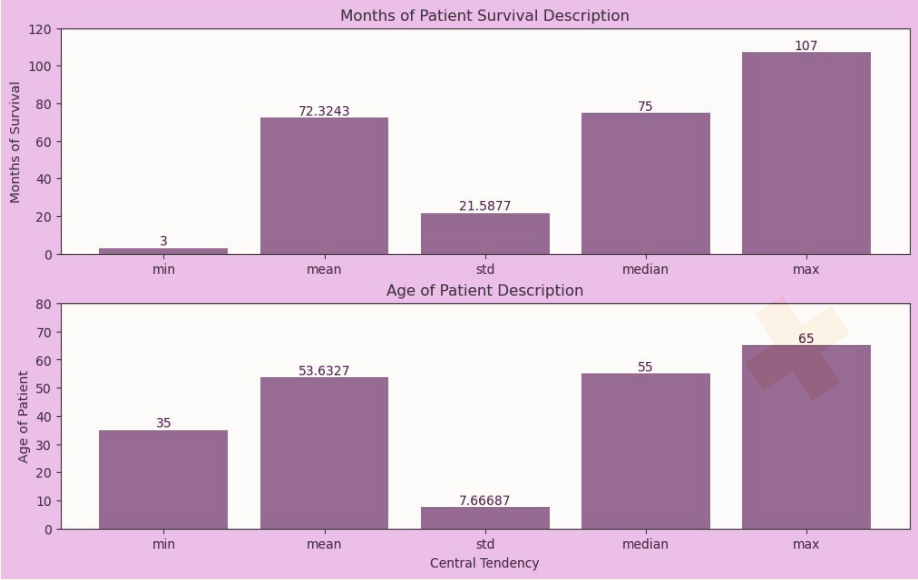
Tasha

Menopause & Survival Months



Results

- As reflected in boxplot, each menopausal **whisker is similar**, including essentially **equal means**
- **Statistically describing** patient's ages & number of months survived **in visual way**

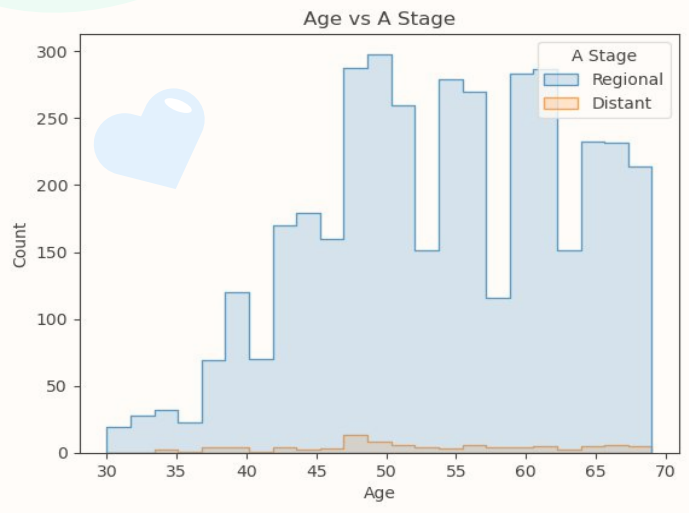


Conclusions

- **Similar means** from boxplot = **similar longevity** for patients, regardless of menopausal status
- **likely no correlation** between menopausal status & months of survival
- **Mean age** of a woman in this study is about **54 years old**

Menopause & Metastasis Sites

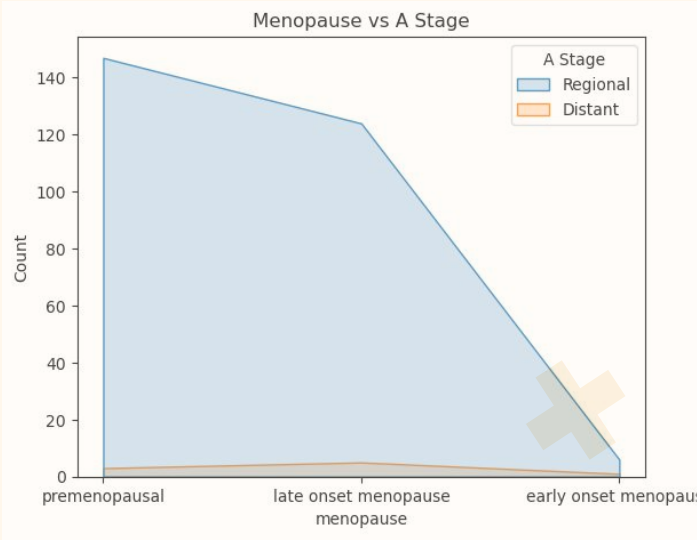
Hidy



Results

In 2% of our patients, neoplasm left the breast & spread into other parts of the body

Regional metastasis sites shows three bursts in data at ages: 45-50, 52-57, & 69-72.



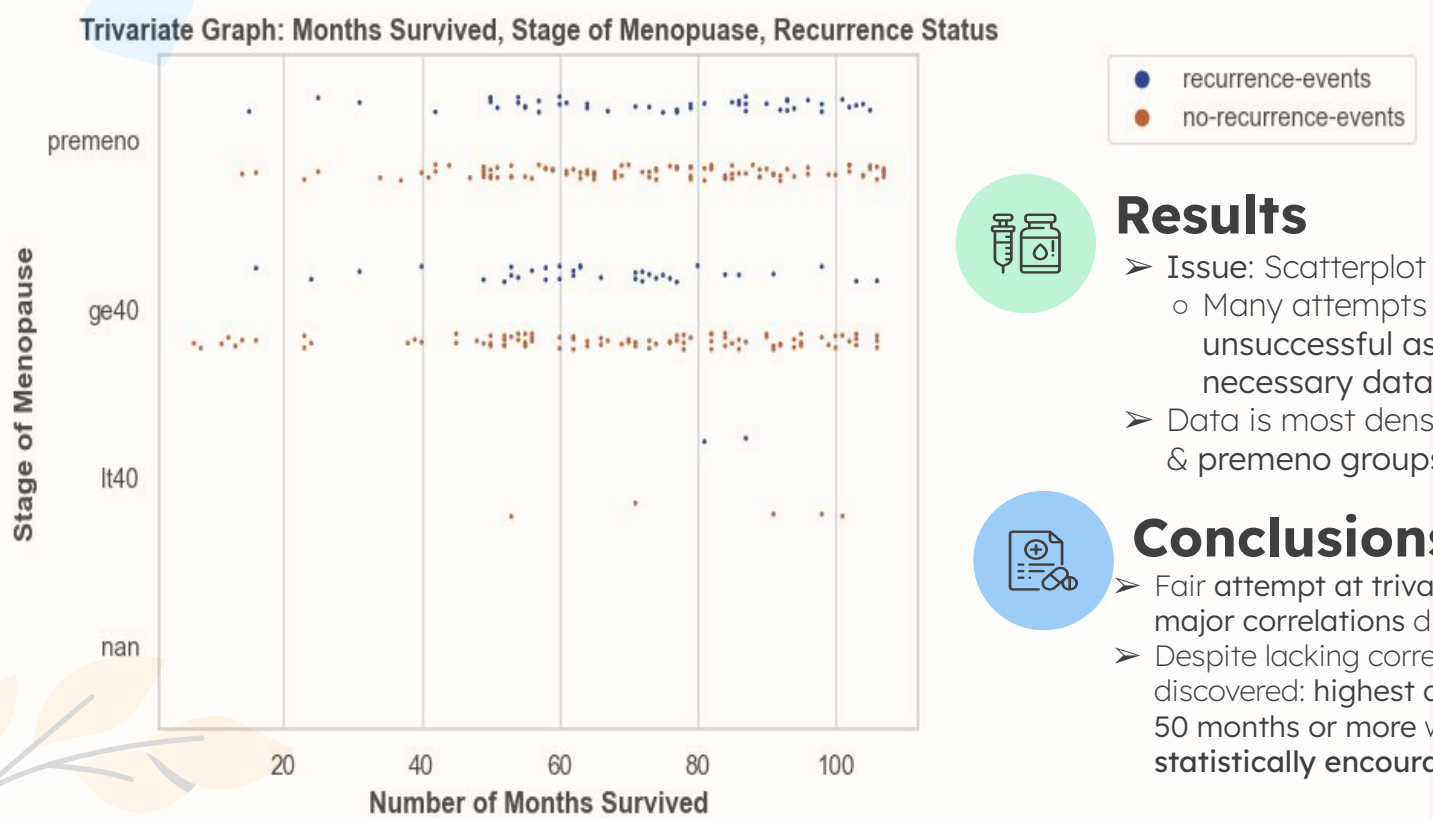
Conclusion

Both illustrations suggest no correlation between the menopausal status & metastasis sites

A peak number of women diagnosed ~ 47 years of age; remains unclear if correlation naturally occurring or are routine medical reminders skewing the data?

Molleigh

Menopause & Trivariate Data



Results

- Issue: Scatterplot with 3 variables shows NaN
  - Many attempts to remove NaN, but unsuccessful as attempts also removed necessary data
- Data is most dense in no-recurrence early onset & premeno groups

Conclusions

- Fair attempt at trivariate data illustration, however, no major correlations discovered visually
- Despite lacking correlations, encouraging insights are discovered: highest data density for women surviving 50 months or more with no-recurrence. This is statistically encouraging news for others

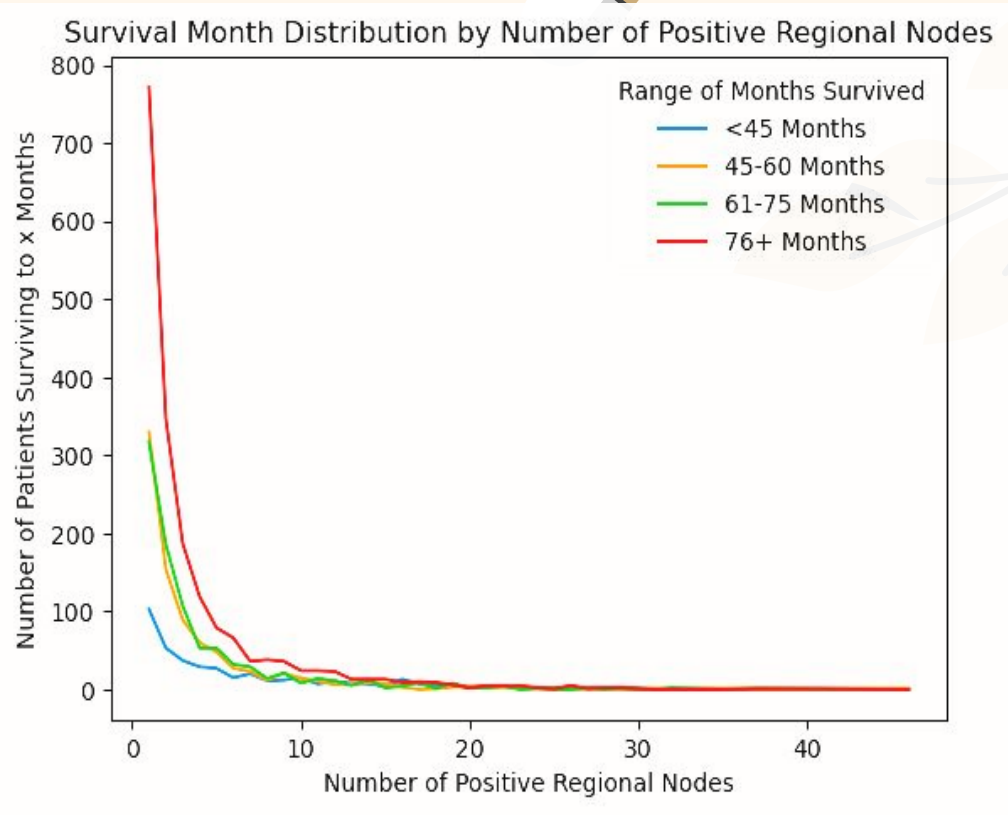
Menopause & Categorical Data

Results

- Until now, we've mostly worked with *continuous* data, but now we're using *categorical* data!
- This **trivariate** illustration is a **great example** of **potential next steps** achievable with this data

Conclusion

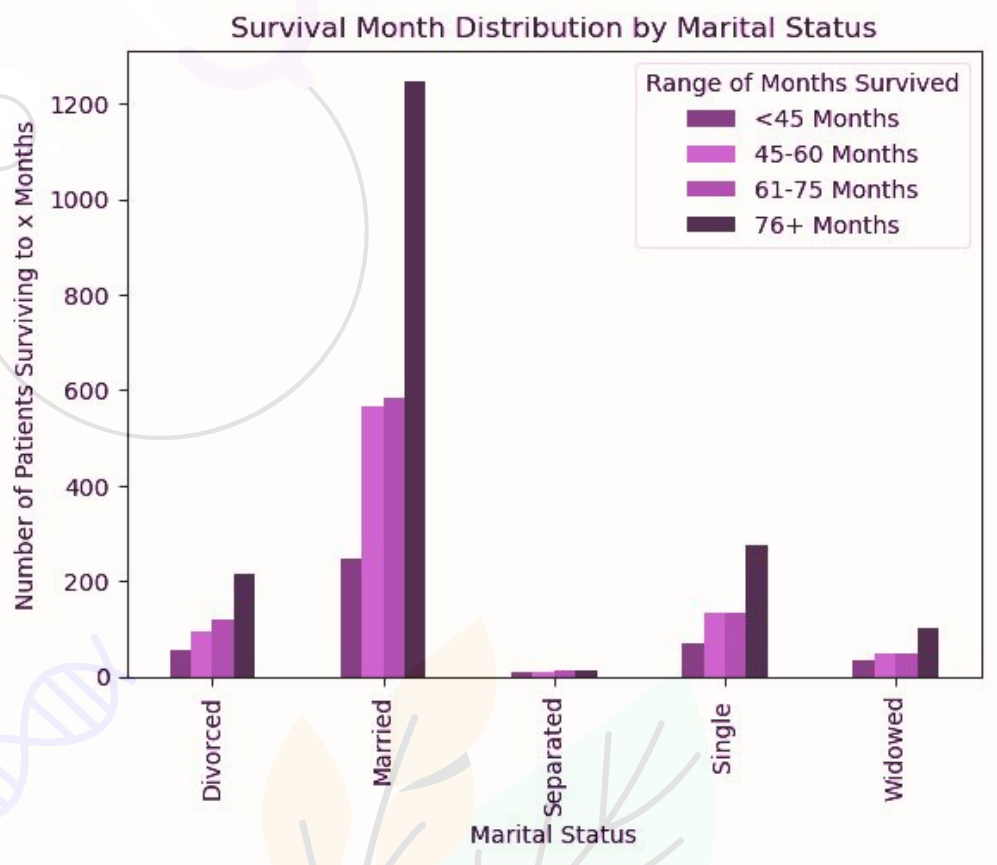
- **Fewer positive regional nodes** (metastasis site) present in the breast **strongly correlates** with the **greatest number of months survived for ALL** the ranges of months survived
  - Women with **less than 10 positive regional nodes** are the **most likely to survive**



Miranda

Menopause & Qualitative Data

Miranda



Results

This **trivariate barchart** shows how **marital status** (in varying age ranges) may affect the **number of patients surviving to *x* months**

Another example of potential next steps

Conclusion

Across all the ranges of months survived, **married patients** are **overwhelmingly most likely** to **survive** breast cancer

Conversely, **separated & widowed patients** are the **least likely** to achieve longevity

04

Potential Next Steps

Potential next steps? Data manipulations we have made? Conclusions to explore?

“Hope is the thing with feathers that perches in the soul, and sings the tune without the words, and never stops at all.”

—Emily Dickinson

Breast Cancer Progression

Potential Next Steps…

In the future, we surmise that **qualitative variables** such as marital status and race, will be the **most expressive and instructive** for data analysis

The **more data analysis** performed on breast cancer research, the **more informed** all women will become about protecting, preventing, and treating breast cancer

Marital Status

Why are married women more likely to survive? Why are separated & widowed women less likely to survive

Positive Regional Nodes

Additional data analysis on positive regional nodes juxtaposed with qualitative variables may be helpful

Recurrence

Recently, Penn Medicine researchers receive $10 million grant for **preventing** breast cancer **recurrence**

Project & Goals

Project

For women with breast cancer, how does her **menopausal status** correlate to her **recurrence** status, **number of months survived**, and **metastasis sites**

Goals

By studying these relationships, we hope to better equip healthcare providers; for example, by **increasing understanding** of breast cancer **risks & causes**, or by **speeding up** breast **cancer predictions & diagnosis.**

Only The Beginning

MSU Data Analysis Bootcamp: Project 1

Stephanie Santiago
Molleigh Hughes
Hidy Vengalil
Tasha Christensen
Miranda Smith