Description of Dataset:

The SEER Program of the NCI did a study on patients with infiltrating duct and lobular carcinoma forms of breast cancer. The patients were diagnosed between 2006 to 2010. The study excluded patients with unknown tumor sizes, patients with lymph nodes along major arteries and the colon, and patients whose survival times were less than 1 month.

Scientific Questions:

Explore different models to find the most significant covariate(s)(e.g. Age, Race, Marital Status, etc.) associated with survival times. Through the process, we will observe the 95% Confidence Interval for the hazard ratio, and answer specific questions related to the covariates such as whether or not older or younger patients have significantly different survival risks. We aim to build a multivariable model that attempts to accurately predict the risks of a patient given specific data for the selected covariates.

Description of the Covariates:

Age - The age in years of the patient at diagnosis of breast cancer.
Race - The race of the patient: black, white, American Indian, Native Alaskan, Asian/Pacific Islanders, & others.
Marital Status - Whether or not the patient is married at the time of diagnosis for breast cancer.
T Stage - The stage of the primary tumor of the breast cancer patient.
N Stage - The stage of how many nearby lymph nodes the breast cancer has spread to.
6th Stage - The classification of the patient's breast cancer based on the size of the tumor and the nodes that the cancer has spread to.
Grade/Differentiate - The description of the removed cancer cells from the breast. A higher grade number (3 = Poorly differentiated) indicates that the cancer is growing quickly and more likely to spread, while a lower grade number (1 = Well differentiated) indicates that the cancer is slowly growing and less likely to spread. Other grades: (2 = Moderately differentiated, 4 = Undifferentiated)
A Stage - Describing where the tumor is: either regional (close to major arteries + colon) or distant from those organs.
Tumor Size - Exact size of the patient's tumor in millimeters.
Estrogen Status - Whether or not the patient has estrogen receptors on the given breast with cancer.
Progesterone Status - Whether or the patient has progesterone receptors on the given breast with cancer.
Survival Months - Number of months the patient survived after being diagnosed with breast cancer.
Status - Whether or not the patient has had an event ("Dead") vs dying after the follow up date ("Alive").

Regional Nodes Examined - Number of regional lymph nodes removed and examined
Regional Nodes Positive - Number of regional lymph nodes found to contain metastases/cancer cells.

Bibliography:

1. JING TENG, January 18, 2019, "SEER Breast Cancer Data", IEEE Dataport, doi: https://dx.doi.org/10.21227/a9qy-ph35.
2. *Breast Cancer.* Breast Cancer. Retrieved April 30th, 2025, from https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

Graph of the Data

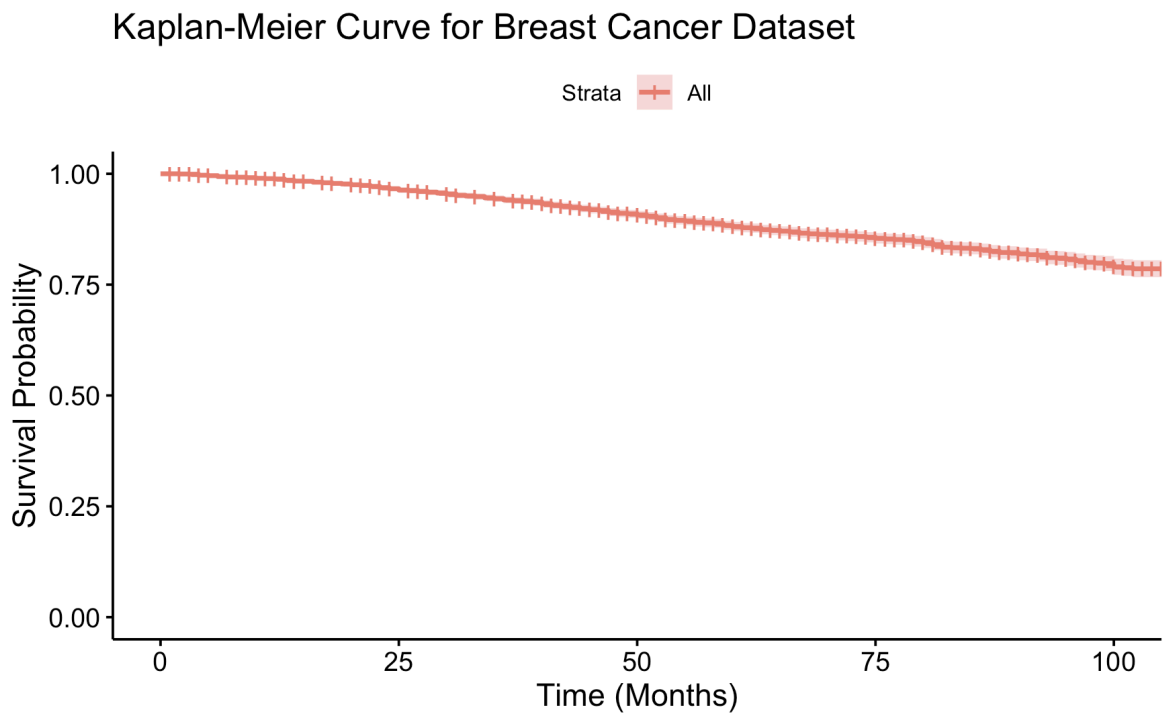
For this analysis, a Kaplan-Meier plot will be created to visualize the survival probability of breast cancer patients over time.

```r
library(survival)
library(survminer)
library(readr)
Breast_Cancer <- read_csv("Breast_Cancer.csv")
head(Breast_Cancer)
Breast_Cancer$Status1 <- ifelse(Breast_Cancer$Status == "Alive", 0, 1)

# Survival Object
breast.surv <- Surv(Breast_Cancer$`Survival Months`, Breast_Cancer$Status1)


breast.fit <- survfit(breast.surv ~ 1)

# Plot the KM curve
ggsurvplot(breast.fit, data = Breast_Cancer,
           xlab = "Time (Months)", ylab = "Survival Probability",
           title = "Kaplan-Meier Curve for Breast Cancer Dataset")
```



Statistical Tools and Complications

- Cox Proportional Hazard Model: We can identify the influence of covariates on survival and help estimate the hazard ratio for each covariate.
- AIC (Akaike Information Criterion):
- Proportional Hazards Assumptions: We can use log-log plots to check the assumptions of proportional hazards.
- Cox ZPH:
- If the assumptions for the cox proportional hazards model are broken such as the log log plot not behaving close enough to parallel curves, then we may need to explore different models such as random forest, ridge regression, lasso regression, etc.

Note for our group:

- Remember to write down in data cleaning that we converted “anaplastic; Grade IV” to “4” for the “Grade” column of the dataset.