Math6450 Survival Analysis and Cox Regression

Ian Tai Ahn

November 10, 2025

Conduct a survival analysis using the following dataset of cancer patients.

Outcome: Overall Survival (Months) - how long the patient survived

Status: Living or Deceased - whether the event occurred

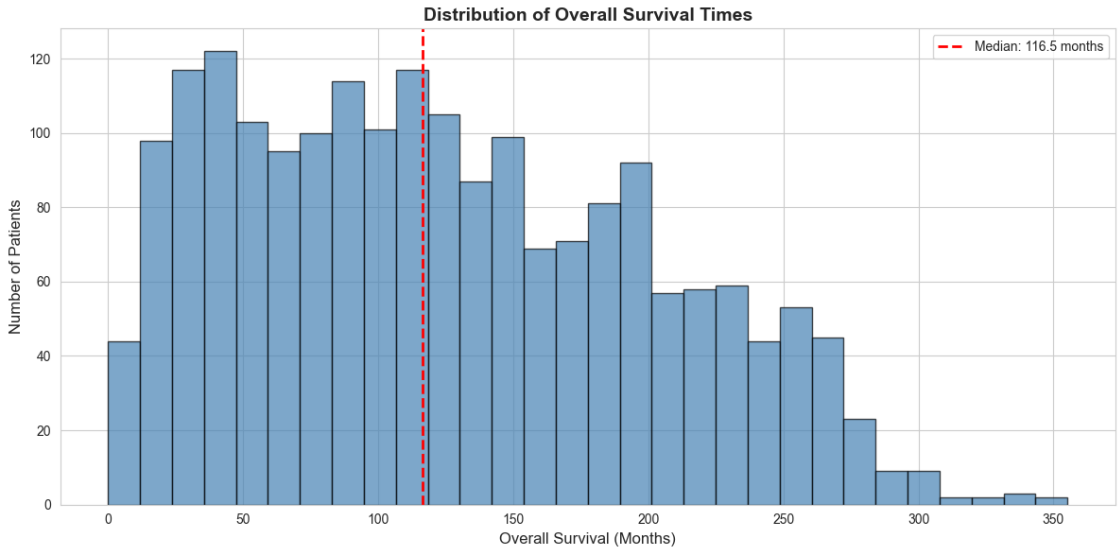
Features: Age, tumor size, cancer stage, treatment types, biomarkers (ER, HER2, PR)

1. Load the dataset and create summary statistics. How many patients are in the study? What is the average age? What percentage of patients survived vs died during the study period?

```
================================================================================
QUESTION 1: DATASET SUMMARY STATISTICS
================================================================================

Total number of patients in the study: 2509
Average age at diagnosis: 60.42 years

Vital Status Distribution:
Patient's Vital Status
Living                  837
Died of Disease         646
Died of Other Causes    497
Name: count, dtype: int64

Survival Summary:
  Survived: 837 patients (33.4%)
  Died: 1672 patients (66.6%)
```

2. Create a histogram showing the distribution of survival times. What do you notice about the shape? Are there any outliers (patients who survived unusually long or short times)?

```
================================================================================
QUESTION 2: DISTRIBUTION OF SURVIVAL TIMES
================================================================================
```

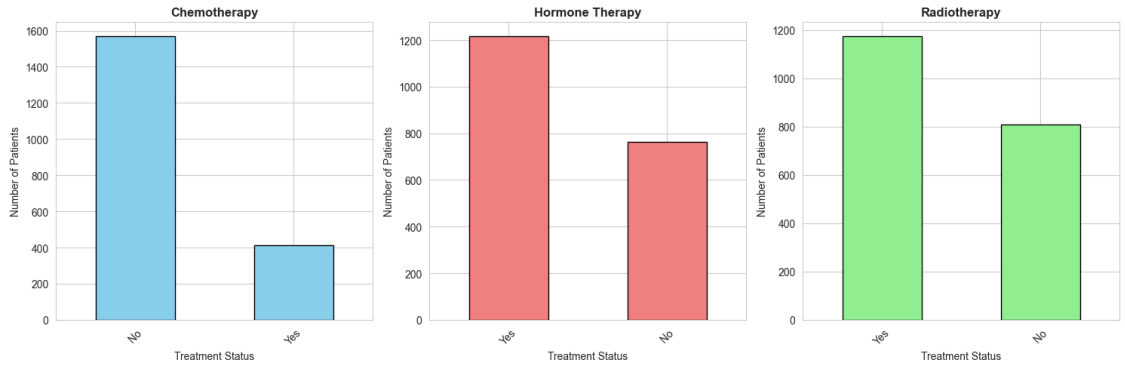



```
Survival Time Statistics:
  Mean: 125.24 months
  Median: 116.47 months
  Min: 0.00 months
  Max: 355.20 months
  Standard Deviation: 76.11 months

Outlier Analysis (using IQR method):
  Number of outliers: 0

Observation: The distribution shows whether survival times are right-skewed
(common in survival data)
```

3. Make a bar chart showing how many patients received each type of treatment (Chemotherapy, Hormone Therapy, Radiotherapy). Which treatment is most common?

```
================================================================================
QUESTION 3: TREATMENT DISTRIBUTION
================================================================================

Treatment Frequencies:

Chemotherapy:
Chemotherapy
No     1568
Yes     412
```

```
Name: count, dtype: int64

Hormone Therapy:
Hormone Therapy
Yes    1216
No      764
Name: count, dtype: int64

Radiotherapy:
Radio Therapy
Yes    1173
No      807
Name: count, dtype: int64
```




```
Most common treatment: Hormone Therapy with 1216 patients
```

4. Create Kaplan-Meier survival curves comparing two groups: Group 1: Patients with ER-Positive tumors Group 2: Patients with ER-Negative tumors

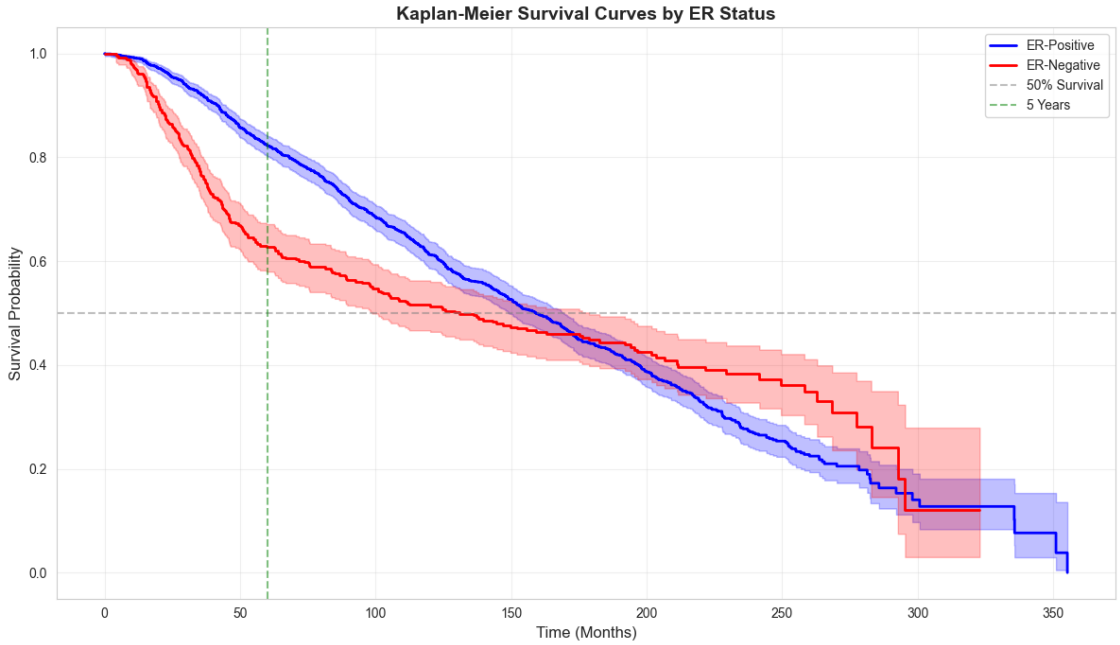
5. Plot both curves on the same graph with a legend.

6. Looking at your plot, which group has better survival? Approximately what percentage of each group is still alive at 5 years (60 months)?

7. What is the median survival time for each group? (The time when 50% of patients have died). If median survival hasn't been reached for a group, explain what that means.

```
================================================================================
QUESTIONS 4-7: KAPLAN-MEIER CURVES BY ER STATUS
================================================================================
```

```
ER-Positive: 1499 patients (after removing NaN)
ER-Negative: 439 patients (after removing NaN)
```

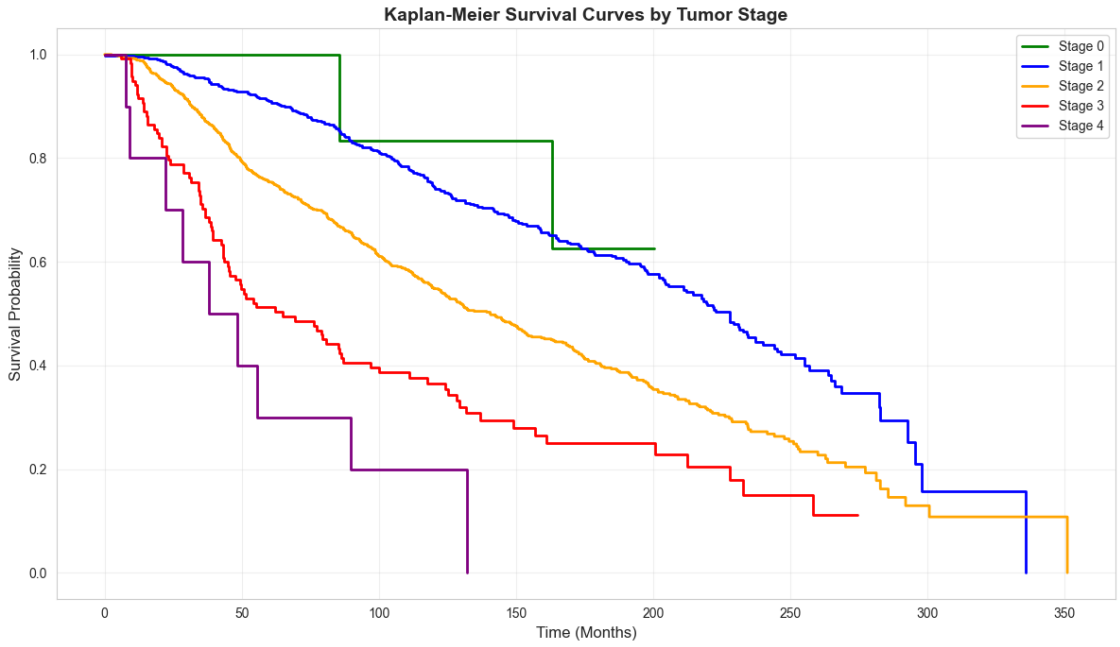



```
5-Year Survival Rates (at 60 months):
  ER-Positive: 82.3%
  ER-Negative: 62.8%

Median Survival Times:
  ER-Positive: 159.23 months
  ER-Negative: 130.87 months

Log-rank test p-value: 0.1102
The difference between groups is not statistically significant ( =0.05)
```

Looks like ER-Positive patients had a longer survival time when looking at the median survival time. However, when looking at the graph they converge around month 175, and then the ER-Negative had a higher chance of survival and then they finally converge again at month 290.

8. Tumor Stage (0, 1, 2, 3, 4) - Create survival curves for each stage on one plot.

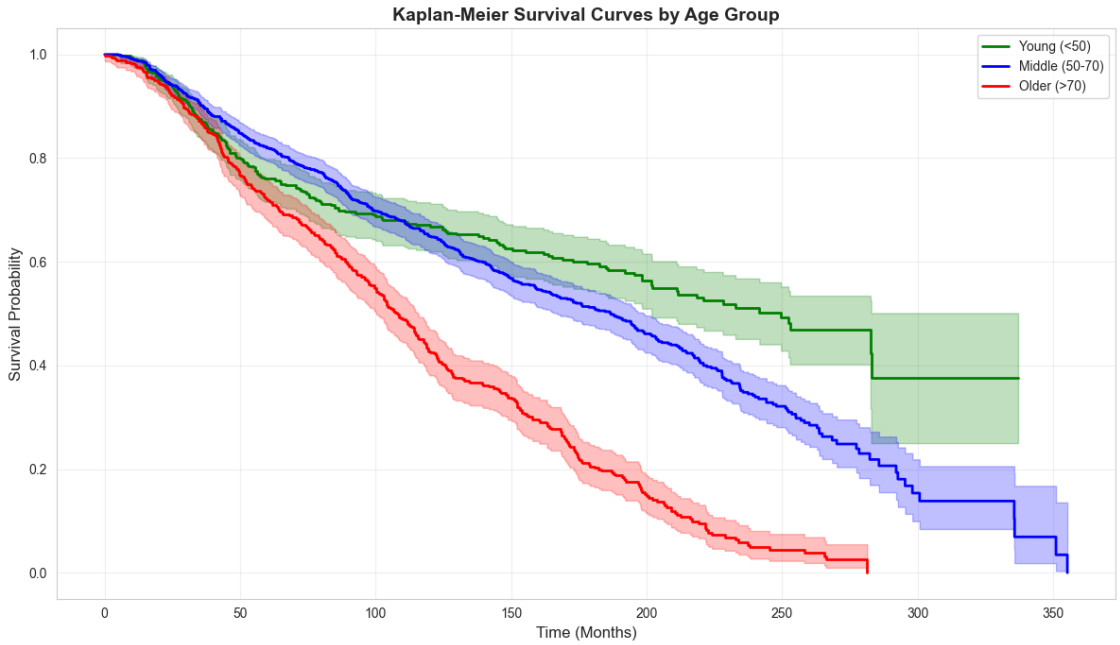
```
================================================================================
QUESTION 8: KAPLAN-MEIER CURVES BY TUMOR STAGE
================================================================================
```




```
Median Survival Times by Stage:
  Stage 0: inf months
  Stage 1: 227.80 months
  Stage 2: 140.60 months
  Stage 3: 64.93 months
  Stage 4: 38.13 months

Pattern: Generally, higher tumor stages show worse survival outcomes.
```

9. Age groups: Create three groups based on age at diagnosis. Young: Under 50 years, Middle: 50-70 years, Older: Over 70 year

```
================================================================================
QUESTION 9: KAPLAN-MEIER CURVES BY AGE GROUP
================================================================================
```




```
Median Survival Times by Age Group:
  Young (<50): 249.53 months
  Middle (50-70): 186.83 months
  Older (>70): 107.77 months

Pattern: Typically, younger patients have better survival outcomes,
while older patients (>70) may have worse outcomes due to age-related factors.
```

10. For both 9 and 10, describe what patterns you see. Which groups have the best and worst survival

```
Looks like stage 0, and young patients have the best rate of survival. Stage 4,
and older patients have the worst survival.
```

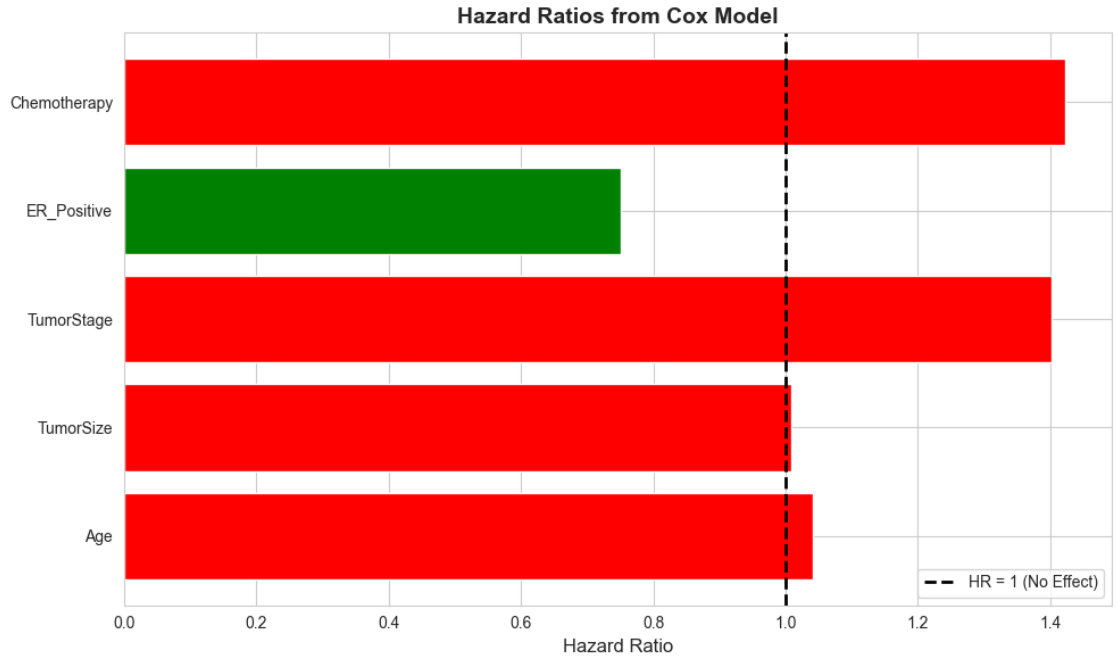
11. Fit a Cox proportional hazards model to predict overall survival using these variables: Age at Diagnosis, Tumor Size, Tumor Stage, ER Status, Chemotherapy

12. Report the hazard ratio for each variable. Which variables have hazard ratios greater than 1 (indicating increased risk)?

```
================================================================================
QUESTIONS 11-12: COX PROPORTIONAL HAZARDS MODEL
================================================================================

Cox Proportional Hazards Model Summary:
```

```
                  coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
Age           0.040972   1.041823  0.003541        0.034032        0.047913
TumorSize     0.008901   1.008941  0.002190        0.004610        0.013193
TumorStage    0.337905   1.402008  0.067047        0.206496        0.469315
ER_Positive  -0.287479   0.750153  0.094749       -0.473183       -0.101774
Chemotherapy  0.352009   1.421921  0.119187        0.118407        0.585611

              exp(coef) lower 95%  exp(coef) upper 95%  cmp to          z  \
covariate
Age                      1.034618             1.049079     0.0  11.570343
TumorSize                1.004620             1.013280     0.0   4.065143
TumorStage               1.229362             1.598898     0.0   5.039829
ER_Positive              0.623016             0.903233     0.0  -3.034112
Chemotherapy             1.125702             1.796087     0.0   2.953418

                         p    -log2(p)
covariate
Age           5.824967e-31  100.437521
TumorSize     4.800309e-05   14.346513
TumorStage    4.659471e-07   21.033330
ER_Positive   2.412452e-03    8.695284
Chemotherapy  3.142763e-03    8.313751

================================================================================
HAZARD RATIOS AND INTERPRETATION:
================================================================================

    Variable  Hazard Ratio  HR 95% CI Lower  HR 95% CI Upper      P-value
         Age      1.041823         1.034617         1.049079 5.824967e-31
   TumorSize      1.008941         1.004620         1.013280 4.800309e-05
  TumorStage      1.402008         1.229359         1.598902 4.659471e-07
 ER_Positive      0.750153         0.623014         0.903237 2.412452e-03
Chemotherapy      1.421921         1.125697         1.796095 3.142763e-03


Interpretation:

Age:
  Hazard Ratio: 1.042
  Effect: INCREASED risk of death
  Significance: statistically significant (p=0.0000)
  Interpretation: Each unit increase in Age increases the hazard by 4.2%

TumorSize:
  Hazard Ratio: 1.009
  Effect: INCREASED risk of death
  Significance: statistically significant (p=0.0000)
```

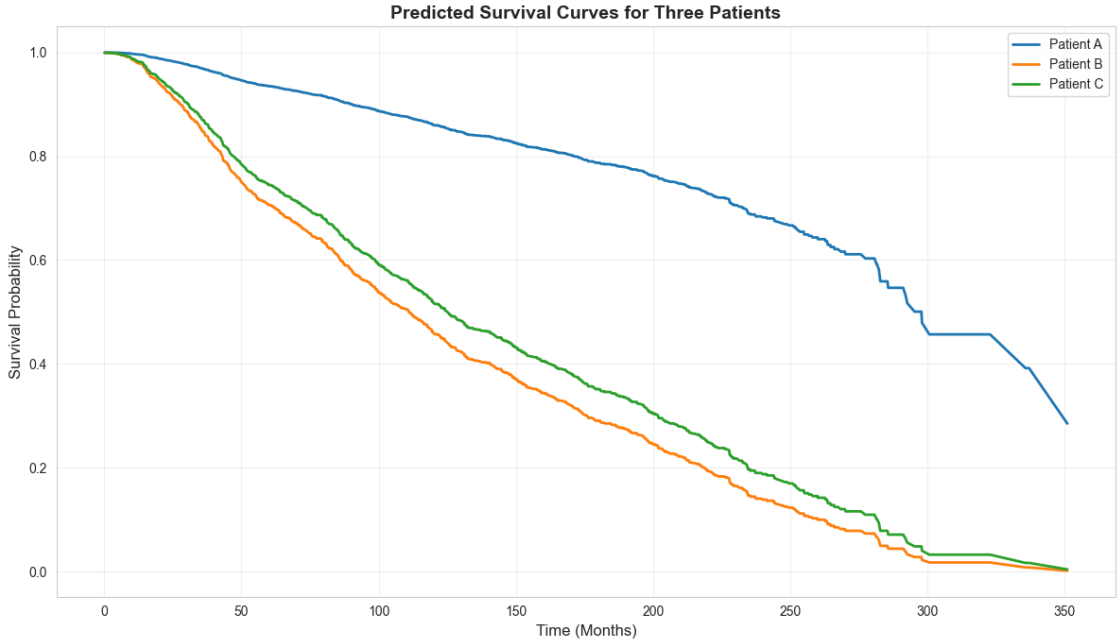
```
  Interpretation: Each unit increase in TumorSize increases the hazard by 0.9%

TumorStage:
  Hazard Ratio: 1.402
  Effect: INCREASED risk of death
  Significance: statistically significant (p=0.0000)
  Interpretation: Each unit increase in TumorStage increases the hazard by 40.2%

ER_Positive:
  Hazard Ratio: 0.750
  Effect: DECREASED risk of death
  Significance: statistically significant (p=0.0024)
  Interpretation: Each unit increase in ER_Positive decreases the hazard by
25.0%

Chemotherapy:
  Hazard Ratio: 1.422
  Effect: INCREASED risk of death
  Significance: statistically significant (p=0.0031)
  Interpretation: Each unit increase in Chemotherapy increases the hazard by
42.2%
```



13. Using your Cox model from Question 12, calculate risk scores for these three patients:

Patient	Age	Tumor Size (mm)	Stage	ER Status	Chemotherapy
Patient A	45	15	1	Positive	No
Patient B	65	30	2	Positive	Yes
Patient C	55	25	2	Negative	Yes

```
================================================================================
QUESTION 13: RISK SCORES FOR THREE PATIENTS
================================================================================

Patient Characteristics:
           Age  TumorSize  TumorStage  ER_Positive  Chemotherapy
Patient A   45         15           1            1             0
Patient B   65         30           2            1             1
Patient C   55         25           2            0             1

================================================================================
RISK SCORES:
================================================================================

Patient A: 0.3228

Patient B: 1.6690

Patient C: 1.4127

Interpretation:
Risk scores are relative to the baseline hazard.
Higher values indicate higher risk of death.

Risk Comparison:
Patient A: Baseline (reference)
Patient B: 5.17x the risk of Patient A
Patient C: 4.38x the risk of Patient A
```


Predicted Survival Curves for Three Patients
Patient A
Patient B
Patient C
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
Time (Months)