Understanding the Data

Variables:

- **Survival Time (Days):** Continuous variable representing how long patients survived after diagnosis.
- **Cancer Type:** Categorical variable with multiple levels (e.g., Stomach, Bronchus, Colon, etc.).
- **Sex:** Categorical variable with two levels (Male, Female).
- **Age:** Continuous variable indicating the patient's age.

Goal: Model the survival time and explore the relationships between survival time and other variables (age, cancer type, sex). This suggests a focus on time-to-event data, making survival analysis a key method to consider.

This is a time-to-event (survival) analysis problem. Therefore, appropriate models include:

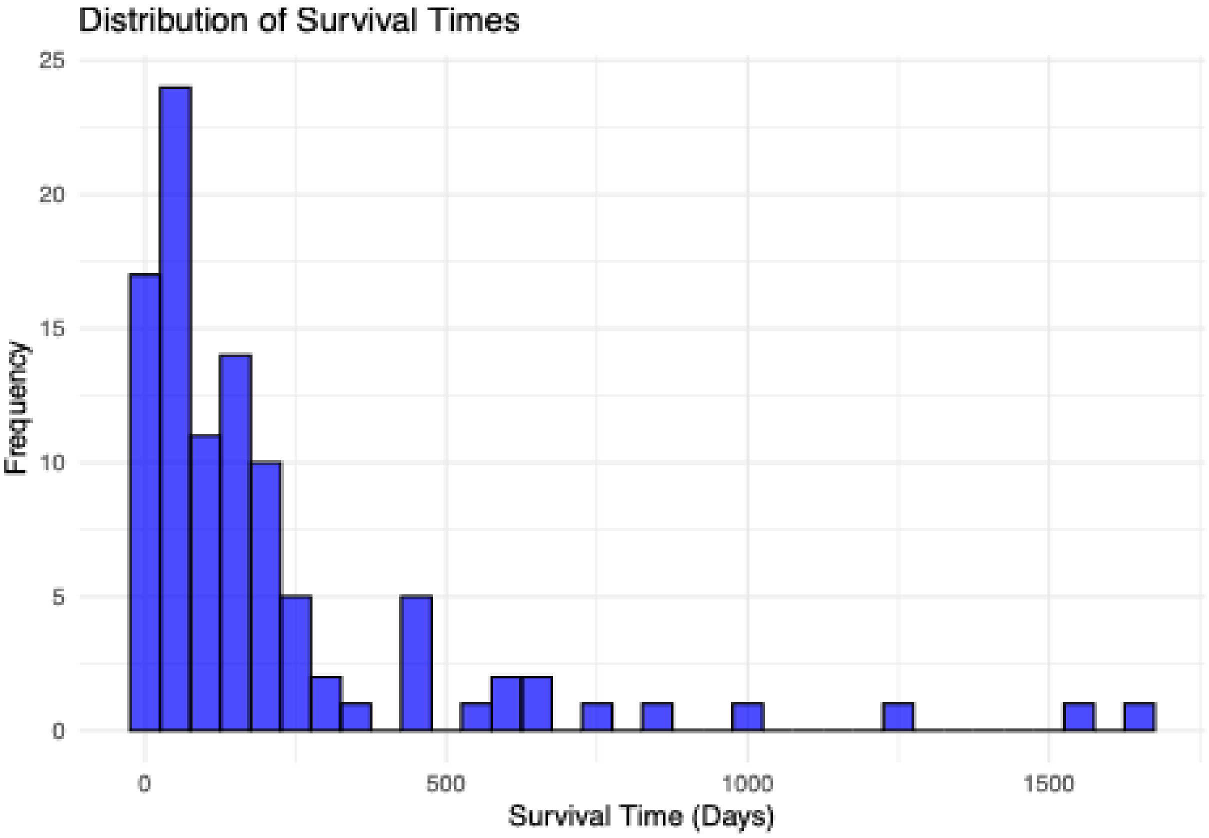
- **Kaplan-Meier Estimator** for non-parametric survival function estimation by cancer type.
- **Log-Rank Test** for comparing survival distributions across groups (e.g., different cancer types).
- **Cox Proportional Hazards Model** to assess the effect of multiple covariates (age, sex, cancer type) on survival time.
- **Weibull, Log or Exponential Models** for parametric survival modelling if certain assumptions are met.

Other Potential Statistical Models

- Parametric Survival Models (e.g., Weibull, Exponential, Log-Normal)
- Accelerated Failure Time (AFT) Model

Visualisations for initial exploration:

1. Distribution of Survival Times:



Analysis:

- The histogram shows a strongly right-skewed distribution of survival times.
- A large proportion of patients have survival times less than 500 days.
- There's a long tail extending beyond 1500 days, indicating some patients survived much longer than others.

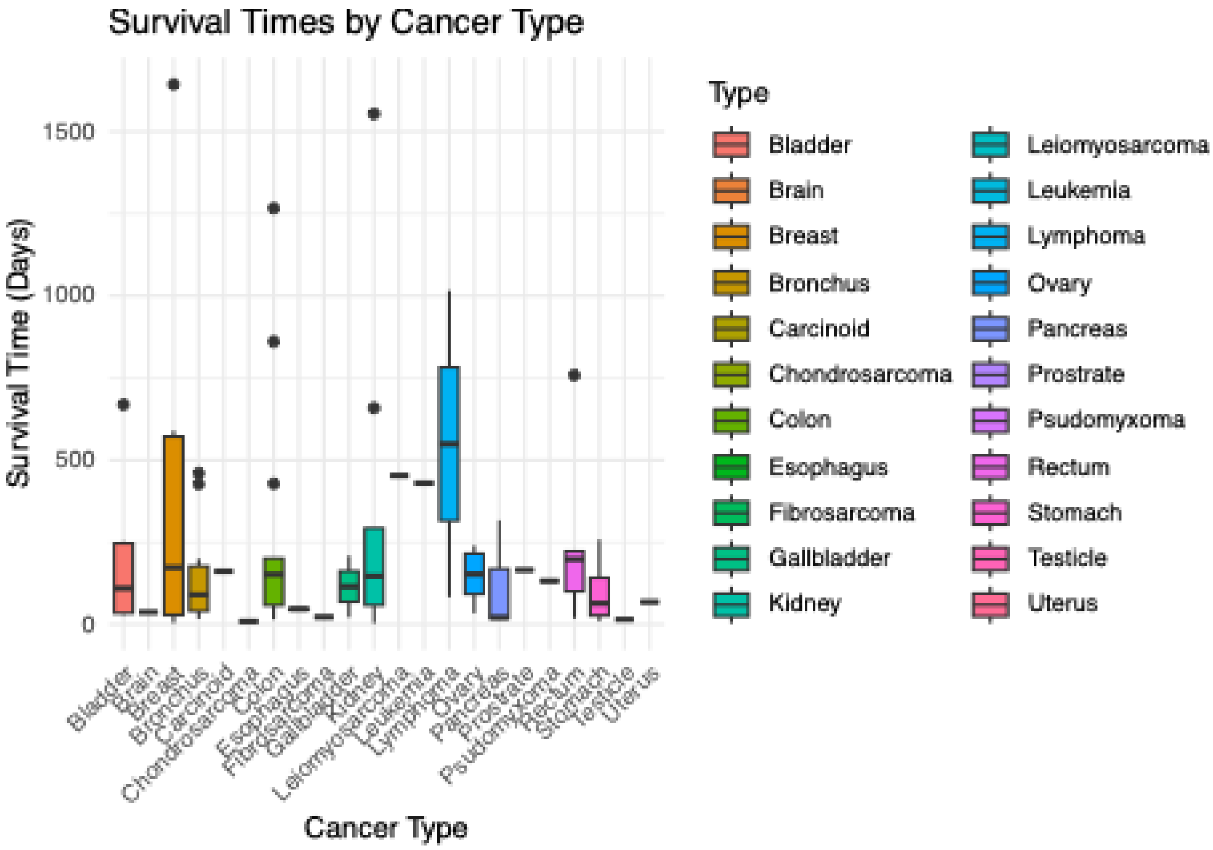
Evaluation:

- This distribution suggests that a non-parametric approach like Kaplan-Meier might be more appropriate than assuming a specific parametric distribution.
- The skewness indicates that median survival time would be a more appropriate measure of central tendency than mean survival time.

Conclusions:

- The variation in survival times is considerable, which warrants investigation into factors that might explain this variability (e.g., cancer type, age, sex).
- The right-skewed nature of the data suggests that while most patients have shorter survival times, there are some long-term survivors who may be of particular interest.

2. Survival Times by Cancer Type:



Analysis:

- There's substantial variation in survival times across different cancer types.
- Some cancer types (e.g., Breast, Colon, Kidney) show wider ranges of survival times compared to others (e.g., Stomach, Ovary).
- Median survival times differ noticeably across cancer types, with some (like Breast) having higher median survival times than others (like Stomach or Bronchus).

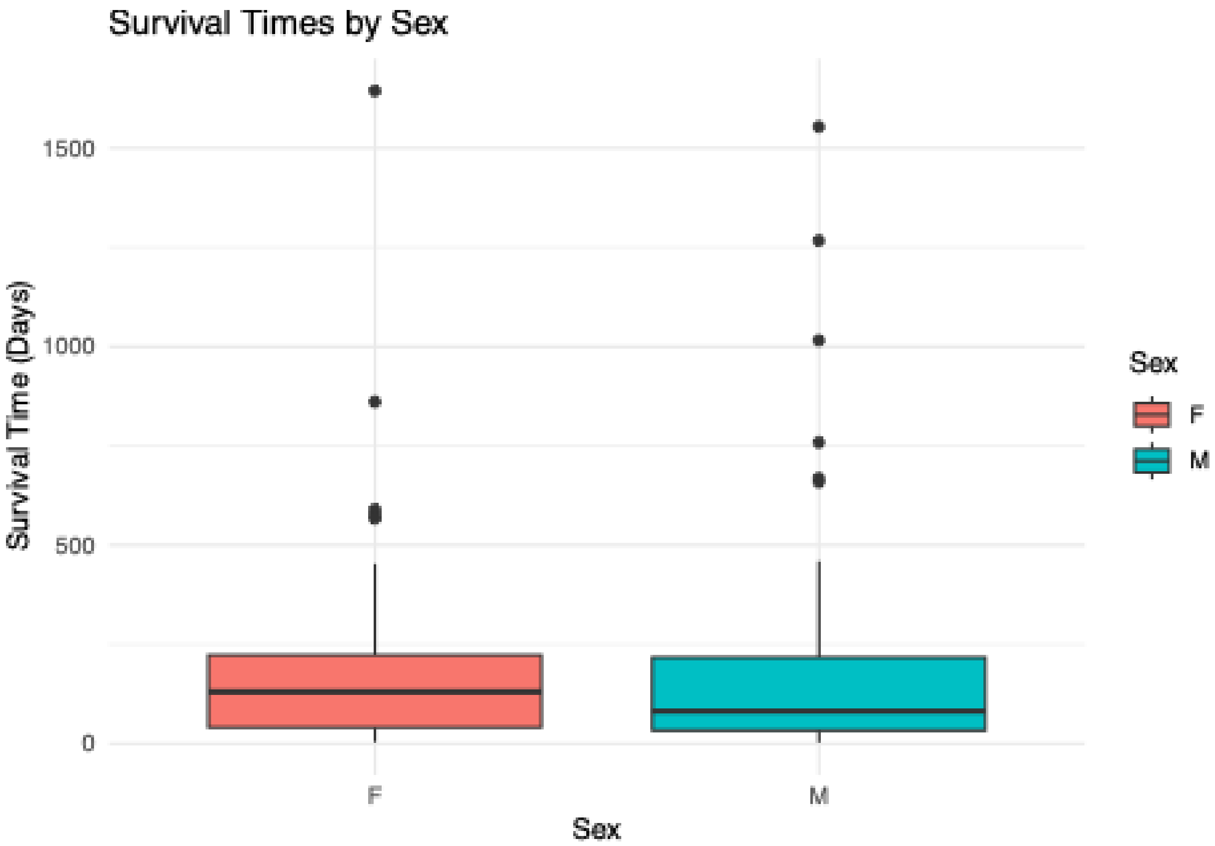
Evaluation:

- The differences in survival distributions suggest that cancer type is likely a significant factor in predicting survival.
- The varying ranges of survival times within each cancer type indicate that other factors (possibly age or sex) may also play a role.

Conclusions:

- Cancer type appears to be an important predictor of survival time and should be included in any predictive model.
- The heterogeneity within cancer types suggests that stratified analysis or inclusion of interaction terms might be necessary in the modelling process.

3. Survival Times by Sex:



Analysis:

- Females appear to have a slightly higher median survival time compared to males.
- The range of survival times for females seems wider than for males, with some extreme high values.
- Both distributions are right-skewed, with some extreme values on the upper end.

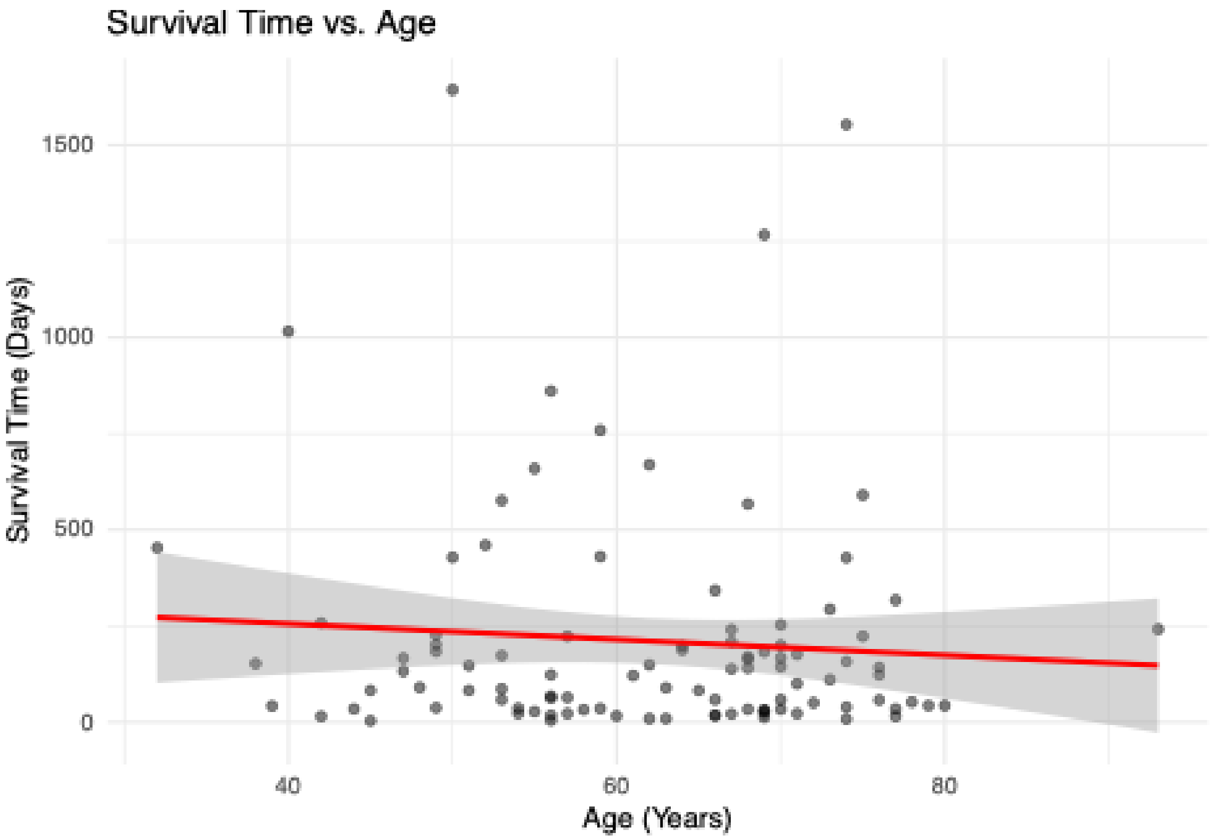
Evaluation:

- While there are differences between sexes, they don't appear as pronounced as the differences between cancer types.
- The overlapping ranges suggest that sex alone may not be a strong predictor of survival time.

Conclusions:

- Sex might be a relevant factor in survival analysis, but its effect may be less significant than cancer type.
- The interaction between sex and other variables (e.g., cancer type, age) should be investigated, as the effect of sex might vary across different subgroups.

4. Survival Time vs. Age:



Analysis:

- There's no clear linear relationship between age and survival time.
- Survival times appear to be highly variable across all age groups.
- There might be a slight tendency for younger patients to have some of the longer survival times, but this is not a strong trend.

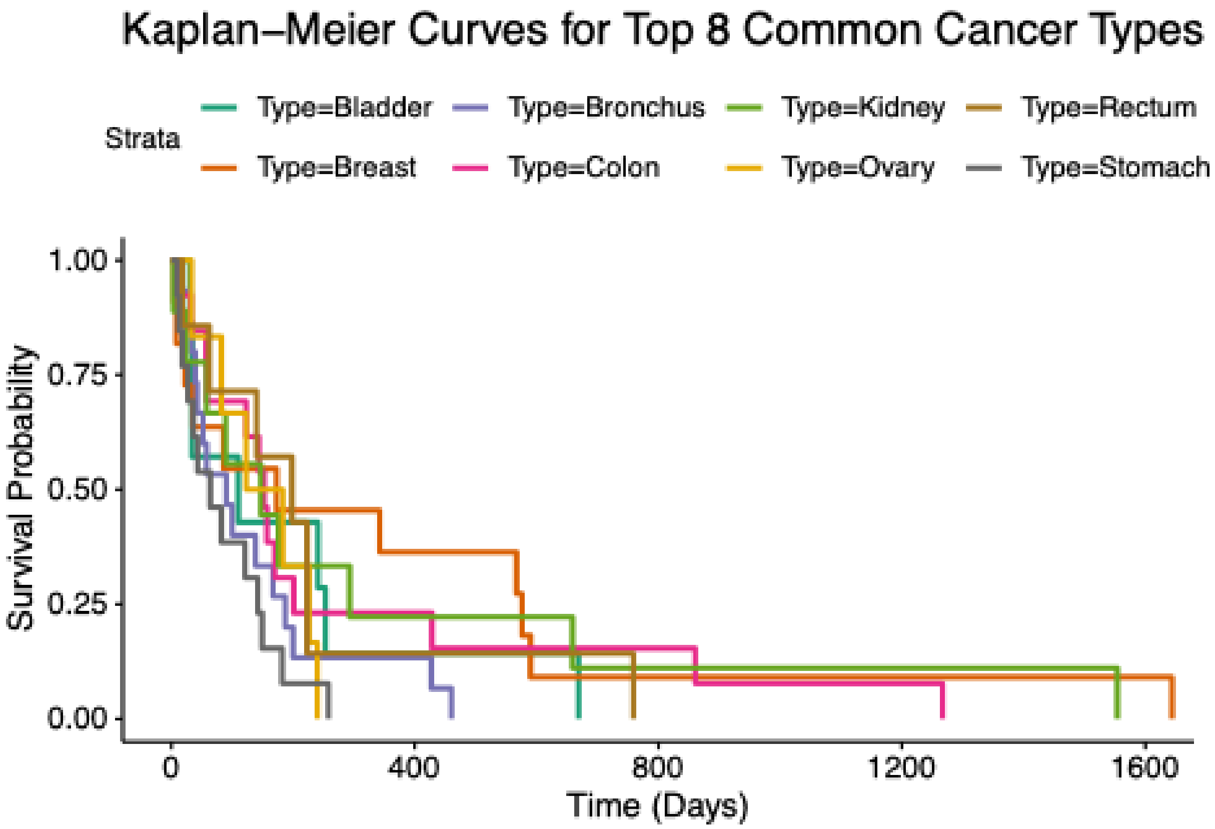
Evaluation:

- The lack of a clear linear relationship suggests that simple linear regression would not be appropriate for modelling the age-survival relationship.
- The high variability indicates that factors other than age are likely important in determining survival time.

Conclusions:

- Age alone does not appear to be a strong predictor of survival time.
- Non-linear relationships or interaction effects between age and other variables should be considered in the modelling process.
- The relationship between age and survival might be better understood when stratified by cancer type or other factors.

5. Kaplan-Meier Curves for Top 8 Common Cancer Types:



Analysis:

- The survival curves for different cancer types show distinct patterns.
- Some cancer types (e.g., Stomach, Bronchus) show steeper declines in survival probability over time compared to others (e.g., Breast, Colon).
- The curves separate early and maintain their relative positions, suggesting persistent differences in survival across cancer types.

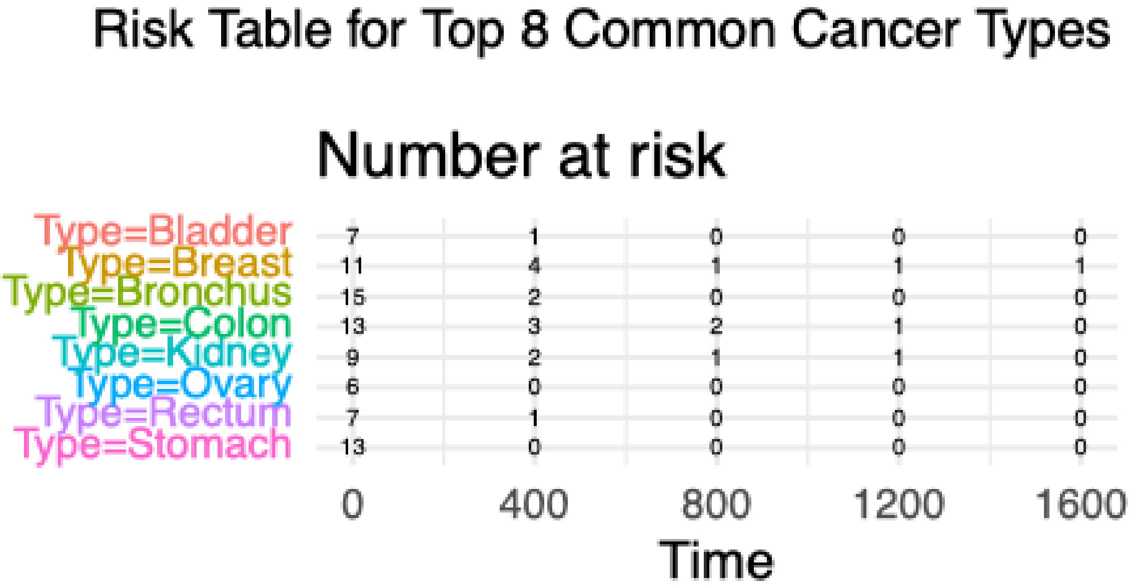
Evaluation:

- The clear separation of curves supports the earlier observation that cancer type is a significant factor in survival.
- The non-crossing nature of most curves suggests that the proportional hazards assumption of a Cox model might be reasonable for these data.

Conclusions:

- Cancer type is a crucial factor in predicting survival and should be a key variable in any survival model.
- The differences in curve shapes suggest that both short-term and long-term survival vary by cancer type.
- A Cox Proportional Hazards model or stratified analysis by cancer type would be appropriate.

6. Risk Table for Top 8 Common Cancer Types:



Risk Table for Top 8 Common Cancer Types

Number at risk

Time	0	400	800	1200	1600
Type=Bladder	7	1	0	0	0
Type=Breast	11	4	1	1	1
Type=Bronchus	15	2	0	0	0
Type=Colon	13	3	2	1	0
Type=Kidney	9	2	1	1	0
Type=Ovary	6	0	0	0	0
Type=Rectum	7	1	0	0	0
Type=Stomach	13	0	0	0	0

Analysis:

- The table shows the number of patients at risk (still alive) at different time points for each cancer type.
- All cancer types show a decrease in the number of patients at risk over time, but at different rates.
- Some cancer types (e.g., Breast) have patients surviving even at the latest time points, while others (e.g., Stomach) have no patients left at risk after a certain point.

Evaluation:

- This table complements the Kaplan-Meier curves by providing exact numbers at risk over time.
- It helps to assess the reliability of the survival probability estimates at different time points.

Conclusions:

- The varying rates of decline in the number at risk support the conclusion that cancer type significantly influences survival patterns.
- For some cancer types, the estimates become less reliable at later time points due to the small number of patients still at risk.
- This information is crucial for interpreting the Kaplan-Meier curves and understanding the long-term survival prospects for different cancer types.

These analyses collectively support the use of survival analysis techniques, particularly those that can account for the differences between cancer types, while also considering the potential influences of age and sex.