General information :

- There are no missing values in this dataset. So there is no need to do data imputation.
- The datatype of 'survival_status_after_5_years' column is integer. It has to be converted to categorical data type.
- The values of 'survival_status_after_5_years' column are not meaningful. Hence they are mapped to 'yes' (survived after 5 years) and 'no' (not survived after 5 years)

Statistics:

- The age of the patients vary from 30 to 83 with the median of 52.
- Although the maximum number of positive lymph nodes observed is 52, nearly 75% of the patients have less than 5 positive lymph nodes and nearly 25% of the patients have no positive lymph nodes
- The dataset contains only a small number of records (306).
- The target column is imbalanced with 73% of values are 'yes'

Univariate plots:

- The number of positive lymph nodes of the survivors is highly dense from 0 to 5. (#5.1)
- Almost 80% of the patients have less than or equal to 5 positive lymph nodes. (#5.2)
- The patients treated after 1966 have the slightly higher chance to survive that the rest. The patients treated before 1959 have the slightly lower chance to survive that the rest. (#5.3 and #5.4)

Multivariate Plots:

By scattering the data points between year_of_treatment and positive_lymph_nodes, we can see the better separation between the two classes than other scatter plots.