Introduction:

The information provided in this document pertains to a breast cancer dataset sourced from Kaggle. This dataset, comprising breast cancer patients, was derived from the November 2017 update of the SEER Program by the NCI, which offers comprehensive cancer statistics based on population data. It specifically focuses on female patients diagnosed with infiltrating duct and lobular carcinoma breast cancer (SEER primary sites recode NOS histology codes 8522/3) during the years 2006-2010. Patients with unspecified tumor size, unexamined regional lymph nodes, positive regional lymph nodes, and those with a survival duration of less than one month were excluded. Consequently, the dataset ultimately includes 4024 patients.

The dataset comprises 16 columns, of which I have utilized 12 for this analysis. These include the following variables: age, race, marital status, T stage, N stage, 6th stage, grade, estrogen status, progesterone status, survival month, and status (whether deceased or alive).

Plot 1:  
  
the figure “age\_cat.png” is a bar graph that shows the distribution of patients across age categories in the dataset. The x-axis represents the age categories, while the y-axis represents the number of patients. The age categories are grouped into ranges (e.g., 0-10, 11-20, etc.).

What the plot shows;

* The highest number of patients falls within the 51-60 age category.
* The number of patients generally appears to increase until the 51-60 age range, then decrease in subsequent age groups.