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.

Plot 2:

This boxplot (age\_race.png) visualizes the distribution of ages at breast cancer diagnosis across different racial groups (White, Black, and Others). The horizontal lines within each box represent the median age (the middle value in the data). The boxes encompass the middle 50% of the data, and the whiskers extend to show the rest of the data points (excluding outliers).

* The boxplot appears to show that Black patients might tend to be diagnosed with breast cancer at younger ages compared to White patients. The median in the "Black" boxplot seems to be lower than the median in the "White" boxplot.
* The spread of ages within each racial group also seems to be different. The box for "Black" patients appears to be taller, suggesting a larger range of ages at diagnosis.

**T-test Results:**

* The p-value between **White vs. Black** (0.0012555792765483553) is statistically **significant**, indicating that there's likely a real difference in the age distribution of breast cancer diagnosis between White and Black patients in this dataset.
* The p-value between White vs. Other (2.8678317568481284e-08) is also highly statistically significant, suggesting a difference between White and Other races as well.
* The p-value between Black vs. Other (0.12798007526770538) is not statistically significant, meaning the evidence is not strong enough to say there's a clear difference in age distribution between Black and Other races in this dataset.

Plot 3:

the bar chart (marital\_Status.png) shows the distribution of patients across marital status categories in a breast cancer study. The green bar represents the number of patients who are married, while the pink bar represents the number of patients who are currently single. The currently single is the add up of all non-married at the time of cancer detection.

The p-value of 0.001 indicates a strong statistical significance. This suggests that the proportion of married women in the dataset is higher than those categorized as 'currently single'.

Plot 4:

This bar chart (Grade\_survival.png) illustrates the average survival months for patients categorized by their breast cancer grade. Each bar represents a grade (Grade 1, Grade 2, Grade 3, and Grade 4), and the height of the bar corresponds to the average number of months patients with that grade survived after diagnosis.

Generally, the average survival months appear to decrease as the breast cancer grade increases. This suggests a potential correlation between higher grades (more aggressive cancers) and shorter survival times.

Plot 5:

The boxes in the plot represent the distribution of survival months for each breast cancer grade. The horizontal line within each box indicates the median survival time.

This box plot () suggests a positive correlation between breast cancer grade and survival months. Patients with lower grades (earlier stages) tend to have higher median survival times compared to patients with higher grades (more advanced stages).

The p-value of the Kruskal-Wallis H-test (0.007483500369062529) also supports this conclusion.

Plot 6:

This box plot shows a relationship between estrogen receptor (ER) status and survival months in breast cancer patients.

"Estrogen Status" refers to whether the breast cancer cells have receptors that can bind to estrogen. Estrogen is a hormone that plays a significant role in the growth and development of breast tissue.

The median survival month appears to be higher for the ER-positive group and the spread of the data (represented by the whiskers) seems to be larger for the ER-negative group (wider box and longer whiskers).

Plot 7:

Similar to the estrogen status plot, this box plot shows the distribution of survival months in breast cancer patients categorized by their progesterone receptor (PR) status.

Similar to ER, **median survival might be longer for the hormone** progesterone **positive groups.**

**Plot 8:**

This violin plot shows the distribution of survival months across different tumor stages for breast cancer patients.

Tumor stage refers to the extent of cancer spread within the body. Typically, lower stages (T1, T2) indicate earlier cancer stages where it's localized or hasn't spread much. Higher stages (T3, T4) represent more advanced cancers that might have spread to lymph nodes or other tissues.

The violin-shaped parts of the plot represent the distribution of survival months for each tumor stage (T1, T2, T3, and T4). The wider area of the violin signifies a higher density of data points at that survival month range.

The violin plot suggests a trend: the median survival month (black line) generally decreases as tumor stage increases (T1 to T4). **This indicates a potential correlation between tumor stage and survival**. Patients with earlier tumor stages (lower stages) tend to have higher median survival times compared to patients with later stages (higher stages).

**T-test result**

**T1 vs. T2, T1 vs. T3, and T1 vs. T4:** The p-values for these comparisons are all extremely small (0.0000 and 0.0005), which signifies **very strong statistical significance**. This suggests that the average survival months for patients with T1 tumors are likely different from those with T2, T3, and T4 tumors. There's a very low chance (less than 0.1%) of observing such extreme differences by random chance if there were truly no underlying differences in survival between these stages.

**Plot 9:**

The boxplot displays the distribution of ages among breast cancer patients categorized by the grade of cancer. The boxplot shows a possible correlation between age and grade of breast cancer.

**T-test**Grades 1 and 3, as well as Grades 2 and 3, exhibit significant differences in age. However, there's no significant age difference between Grade 1 and Grade anaplastic; Grade IV, nor between Grade 2 and Grade anaplastic; Grade IV, or between Grade 3 and Grade anaplastic; Grade IV. These findings underscore the importance of considering age variations across cancer grades in treatment planning and prognosis assessment.

Plot 10:

This scatter plot shows the relationship between tumor size and survival months for breast cancer patients.

There appears to be no clear linear relationship between tumor size and survival months. The data points are scattered throughout the plot, and there's no distinct upward or downward trend. This suggests that tumor size might not be a strong predictor of survival months in this dataset.

Plot 11:

This boxplot shows the distribution of survival months for breast cancer patients across different N stages (N1, N2, and N3). N stage in breast cancer refers to lymph node involvement. Higher N stages generally indicate a more advanced cancer stage where the cancer has spread to lymph nodes.

The median survival month appears to decrease as N stage increases (N1 to N3). This suggests a **potential negative correlation between N stage and survival months**. Patients with higher N stages (more lymph node involvement) tend to have lower median survival times compared to patients with lower N stages.

Thesignificant p-value (indicates that the distributions are unlikely to be the same across all stages. In other words, there's a **statistically significant difference in survival months between the N stage groups**.

Plot 12:

This box plot shows the distribution of tumor sizes in breast cancer patients categorized by whether they are alive or dead. The very small p-value (0.0000) from the t-test indicates a statistically significant difference in tumor sizes between the two groups.

The p-value of 0.0000 signifies a very strong statistical significance. This means it's highly unlikely (less than 0.1%) that we would observe such a difference in average tumor sizes between the two groups if there were truly no underlying difference in reality.

Plot13: