Title: Project 2

Author: Tamiru Workneh

Date: 2024-10-09

Introduction and Problem Definition

I used the Breast Cancer Data Set for this project. This data is obtained from National Cancer Institute (NCI)-Surveillance, Epidemiology, and End Results (SEER) program Breast cancer is one of the most common diseases in the world and mainly affects women, but can also affect men. Family history and other factors such as drinking alcohol, obesity, and mutation in certain genes may increase breast cancer. Some symptoms include a lump or thickening in or near the breast, change in size, scaly or swollen skin on the nipple, etc. Several tests used to diagnose breast cancer are clinical breast exams, mammograms, MRIs, ultrasound exams, and biopsy. If the test shows the prevalence of breast cancer, the decision about the treatment of breast cancer is made based on the size of the cancer how quickly the cancer cell is growing, how the treatment works, or how likely the cancer may come back. This particular data set has 4024 entries and 16 variables. Some variables include different stages of breast cancer such as N Stage (the extent to which the cancer has spread to the lymph nodes), T stage (different stages of tumors), and grade which indicates whether the cancer cells are differentiated or still resemble normal cells.

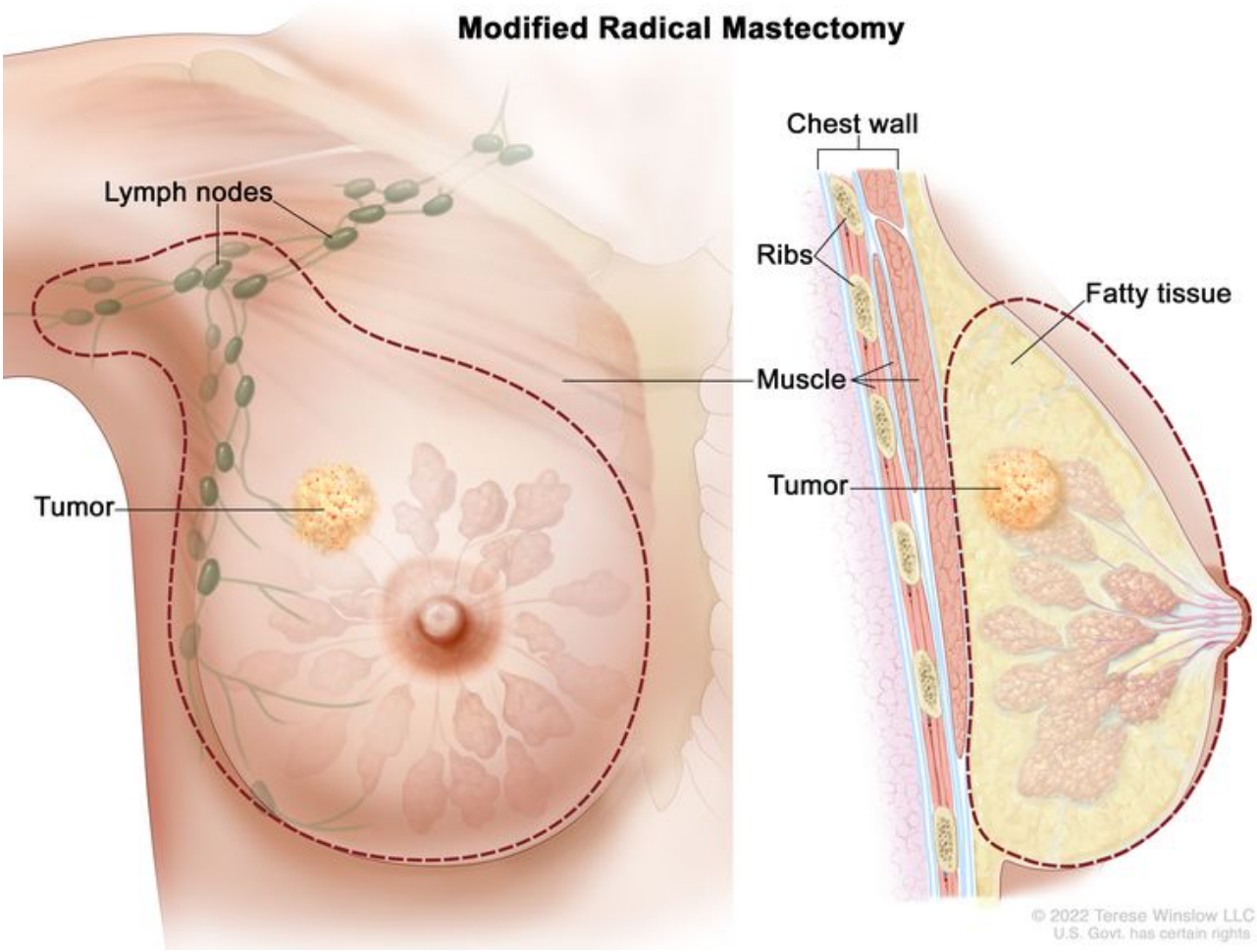
Question: Based on the given data, can we predict the survival rate of breast cancer based on age, Tumor size, stage of cancer (T stage), N Stage, estrogen, and progesterone statuses, and whether or not the regional nodes are positive?

Anatomy of Breast Cancer

The image below illustrates a breast with cancer, indicated by the red dashed line, highlighting the area requiring surgical removal.

In [437…
```
from IPython.display import Image, display
image_path = '/Users/tworkneh/Downloads/415523-750.jpg' #importing the anato
```

```
display(Image(filename=image_path))
```



Importing our data and overview of the imported data.

In [402…

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('Breast_Cancer.csv')
print(df.head(3)) #shows the first three rows (python has 0 index)
print(df.columns)
print(df.dtypes) #data type
```

```
   Age   Race Marital Status T Stage  N Stage 6th Stage  \
0   68  White        Married       T1      N1       IIA
1   50  White        Married       T2      N2      IIIA
2   58  White       Divorced       T3      N3      IIIC

               differentiate Grade   A Stage  Tumor Size Estrogen Status  \
0      Poorly differentiated     3  Regional           4        Positive
1  Moderately differentiated     2  Regional          35        Positive
2  Moderately differentiated     2  Regional          63        Positive

  Progesterone Status  Regional Node Examined  Reginol Node Positive  \
0            Positive                      24                      1
1            Positive                      14                      5
2            Positive                      14                      7

   Survival Months Status
0               60  Alive
1               62  Alive
2               75  Alive
Index(['Age', 'Race', 'Marital Status', 'T Stage ', 'N Stage', '6th Stage',
       'differentiate', 'Grade', 'A Stage', 'Tumor Size', 'Estrogen Status',
       'Progesterone Status', 'Regional Node Examined',
       'Reginol Node Positive', 'Survival Months', 'Status'],
      dtype='object')
Age                       int64
Race                     object
Marital Status           object
T Stage                  object
N Stage                  object
6th Stage                object
differentiate            object
Grade                    object
A Stage                  object
Tumor Size                int64
Estrogen Status          object
Progesterone Status      object
Regional Node Examined    int64
Reginol Node Positive     int64
Survival Months           int64
Status                   object
dtype: object
```

Handling Missing Values

The following code shows the missing values and if there is no missing value, our data is complete and ready for analysis.

In [139…
```
print(df.isnull().sum())
```

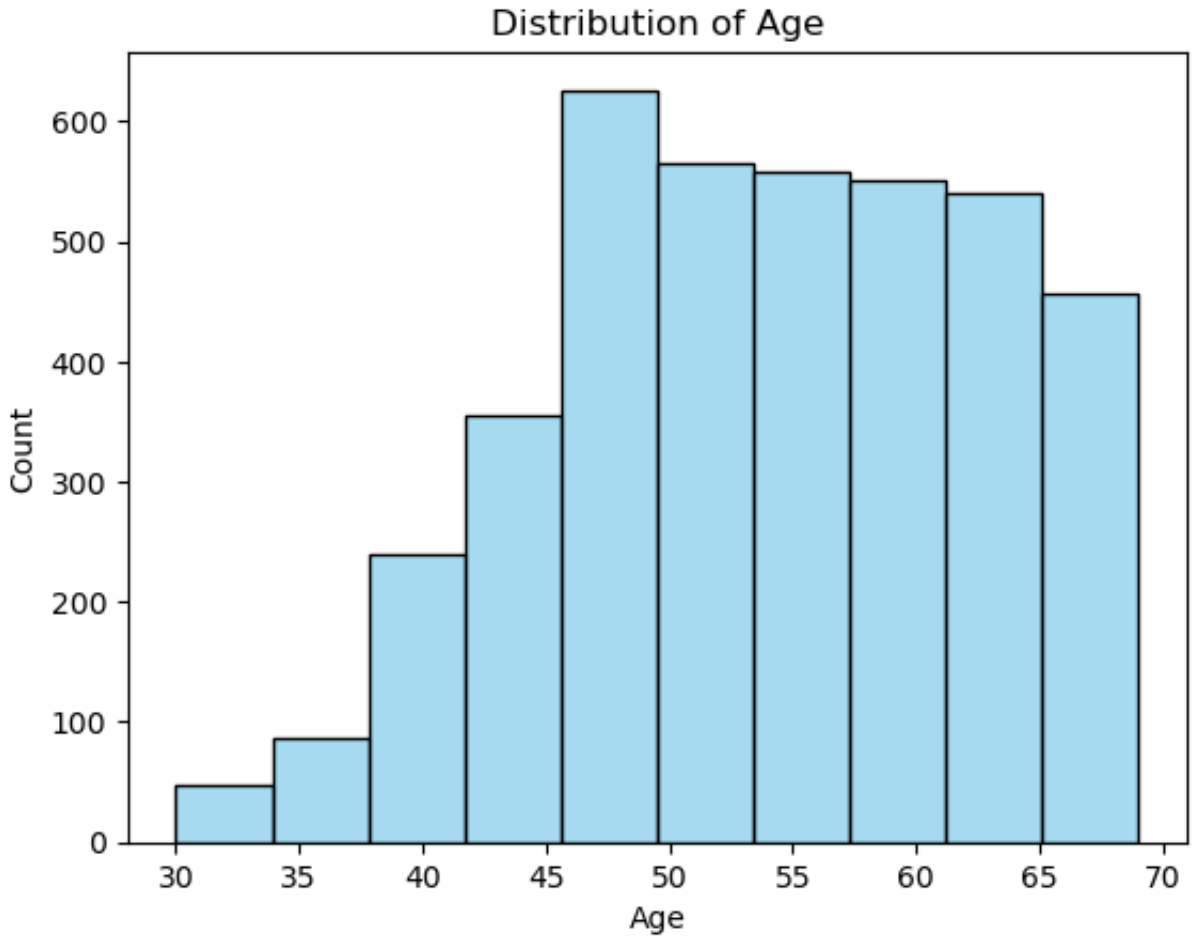
```
Age                        0
Race                       0
Marital Status             0
T Stage                    0
N Stage                    0
6th Stage                  0
differentiate              0
Grade                      0
A Stage                    0
Tumor Size                 0
Estrogen Status            0
Progesterone Status        0
Regional Node Examined     0
Reginol Node Positive      0
Survival Months            0
Status                     0
dtype: int64
```

The Demographics of Breast Cancer Patients (Age, Race, and Marital Status)

Age is considered to be one of the significant factors in breast cancer. The following data shows a substantial increase in breast cancer with age, with the highest counts around age 46 to 55. The data suggests the significance of age and how yearly checkups can be beneficial for women in their 40s and 50s.

In [601…

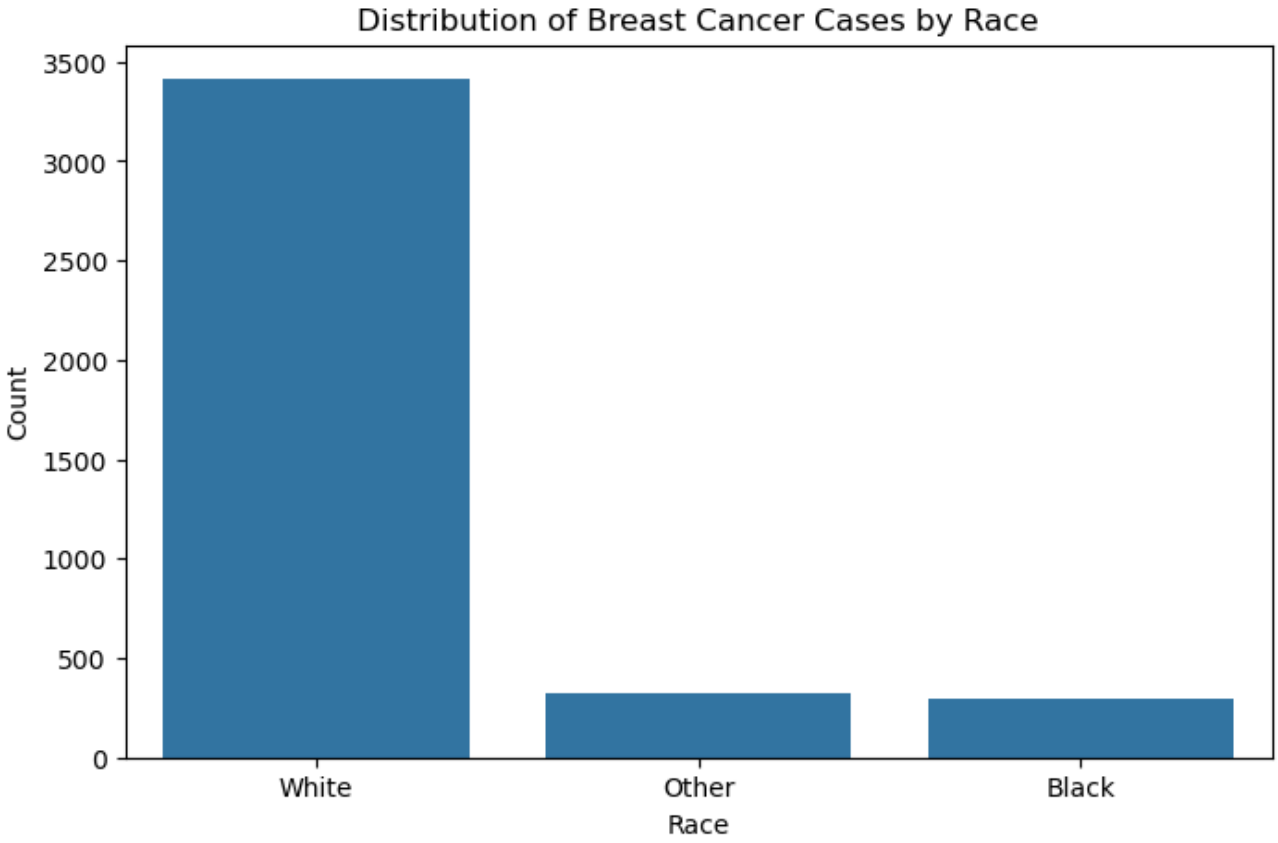
```
#Age
age =df.Age.unique()
#print(age)
sns.histplot(df['Age'],color ='skyblue', bins = 10)
plt.title("Distribution of Age")
plt.show()
print(df.Age.mean())
print(df.Age.median())
print(df.Age.mode())
```




```
53.97216699801193
54.0
0    46
Name: Age, dtype: int64
```

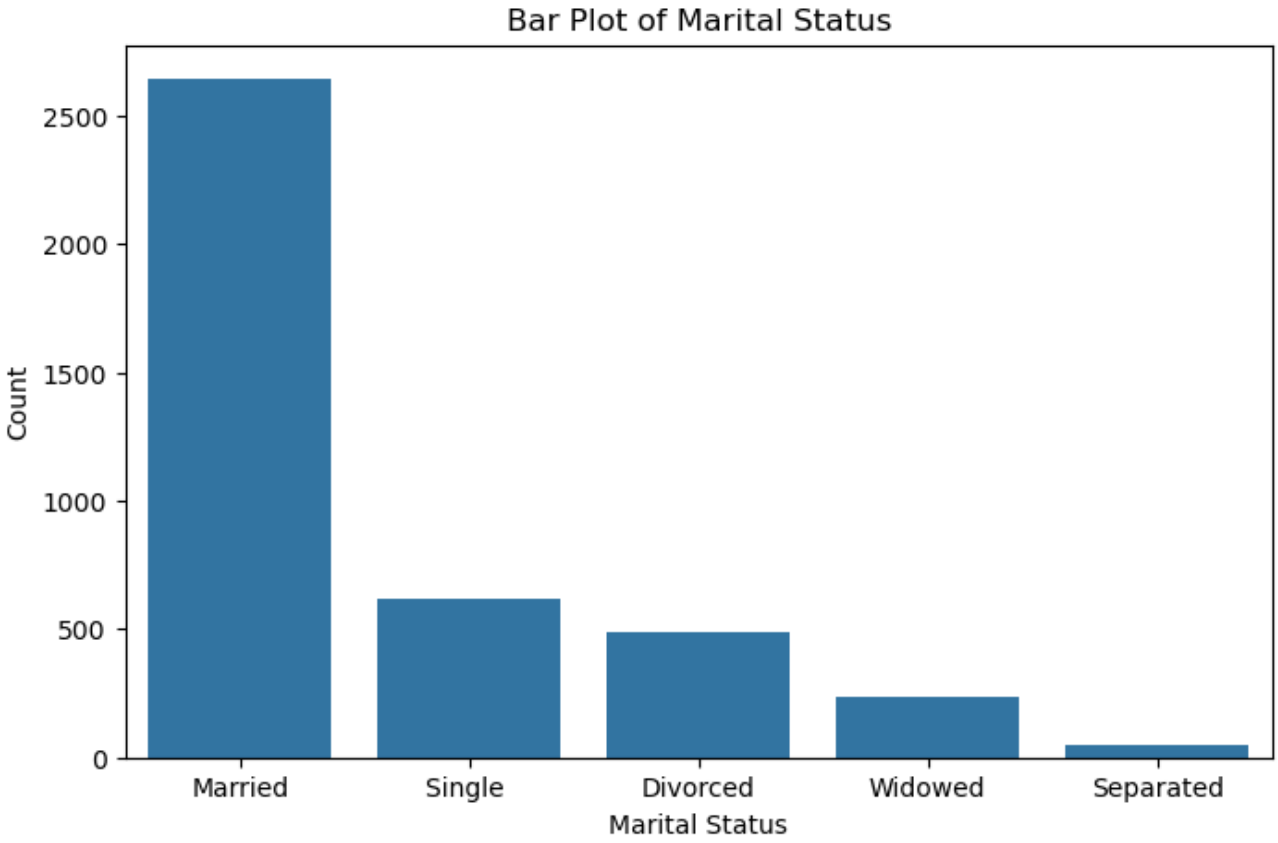
In [751…

```
#Race
race = df.Race.value_counts()
frame = pd.DataFrame(race)
#print(frame)
plt.figure(figsize=(8, 5))
sns.barplot(frame, x='Race', y= df.Race.value_counts())
plt.title('Distribution of Breast Cancer Cases by Race')
plt.xlabel('Race')
plt.ylabel('Count')
plt.show()
```



In [757…

```
#Marital Status
marital_status = df['Marital Status'].value_counts()
ms = pd.DataFrame(marital_status)
plt.figure(figsize=(8,5))
sns.barplot(ms, x='Marital Status', y=df['Marital Status'].value_counts())
plt.title('Bar Plot of Marital Status')
plt.xlabel("Marital Status")
plt.ylabel('Count')
plt.show()
```



Categorical Variables

Most of the columns are categorical variables and I am going to separate categirical variables from numerical variables.

In [732…
```
#df.rename(columns = {'TStage' : 'T Stage'}, inplace=True)

df_encoded = pd.get_dummies(df, columns=['Race', 'Marital Status', 'T Stage
#print(df_encoded)
```

Normalizing Numerical Variables

I separated the categorical variables since most of the columns are categorical. Now, I'm going to normalize the columns like 'Age', 'Tumor Size', Regional Node Examined', 'Reginol Node Positive', and 'Survival Months'. Normalization is a critical step in data analysis. It is important in improving data interpretation, reducing the impact of outliers by scaling the data, etc.

In [406…
```
numerical_cols = ['Age', 'Tumor Size', 'Survival Months', 'Regional Node Exa
scaler = MinMaxScaler()
```

```
df_encoded[numerical_cols] = scaler.fit_transform(df_encoded[numerical_cols]
print("Normalized DataFrame head:")
print(df_encoded.head(3))
```

```
Normalized DataFrame head:
        Age 6th Stage Grade   A Stage  Tumor Size  Regional Node Examined  \
0  0.974359       IIA     3  Regional    0.021583                0.383333
1  0.512821      IIIA     2  Regional    0.244604                0.216667
2  0.717949      IIIC     2  Regional    0.446043                0.216667

   Reginol Node Positive  Survival Months  Race_Other  Race_White  ...  \
0               0.000000         0.556604       False        True  ...
1               0.088889         0.575472       False        True  ...
2               0.133333         0.698113       False        True  ...

   T Stage _T3  T Stage _T4  N Stage_N2  N Stage_N3  \
0        False        False       False       False
1        False        False        True       False
2         True        False       False        True

   differentiate_Poorly differentiated  differentiate_Undifferentiated  \
0                                 True                           False
1                                False                           False
2                                False                           False

   differentiate_Well differentiated  Estrogen Status_Positive  \
0                              False                      True
1                              False                      True
2                              False                      True

   Progesterone Status_Positive  Status_Dead
0                          True        False
1                          True        False
2                          True        False

[3 rows x 25 columns]
```

Distribution of T Stage

The following graph displays the T Stage (tumor stage) representing the size and extent of cancer cells in the breast, ranging from T1 to T4. T1 indicates the smallest size (less than 2cm), while T4 indicates the largest size and a more advanced spread of the disease. In addition, T3 indicated the spread to the nearby tissues, but not distant tissues, however, T4 is the highest stage in cancer indicating the spread to the nearest organs such as the chest wall and skin. More than 540 people have tumor larger than 50 millimeters, but only 10 of them have tumor grown into the chestwall and skin.

In [264…

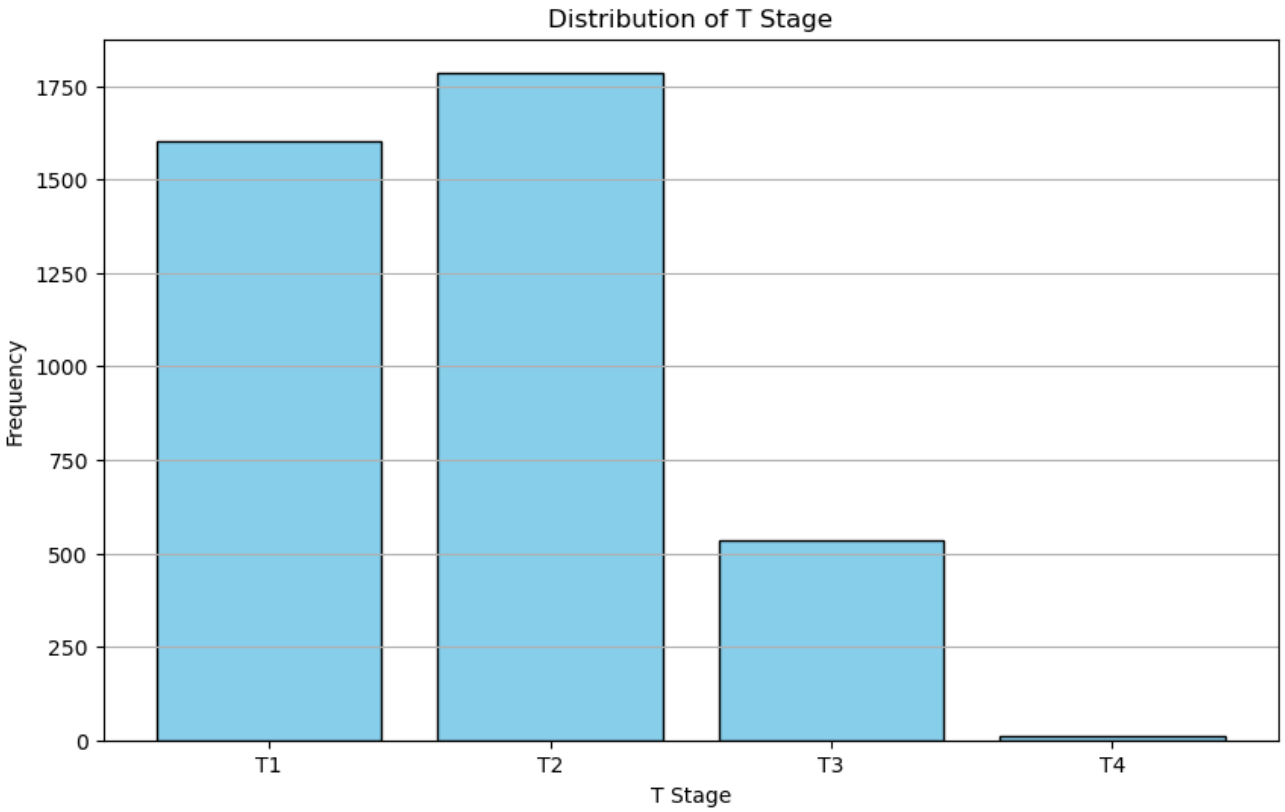
```
import pandas as pd


t_stage_counts = {
    'T1': 1603,
    'T2': 1786,
    'T3': 533,
    'T4': 10
}
df_counts = pd.DataFrame(list(t_stage_counts.items()), columns=['T Stage', '
print(df_counts)

#Plotting the distribution of T Stage
plt.figure(figsize=(10, 6))
plt.bar(df_counts['T Stage'], df_counts['Frequency'], color='skyblue', edgec
plt.title('Distribution of T Stage')
plt.xlabel('T Stage')
plt.ylabel('Frequency')
plt.grid(axis='y')

# Show the plot
plt.show()
```

```
  T Stage  Frequency
0      T1       1603
1      T2       1786
2      T3        533
3      T4         10
```



Distribution of N Stage

The N Stage is a standardized way to describe the extent of cancer in patients. N1 indicates less spread to the lymph nodes, while N3 shows excessive lymph node involvement. Generally, higher lymph node involvement suggests a greater chance for the cancer cells to spread throughout the body. However, according to the graph, N1 has significantly higher levels than N2 and N3, indicating less involvement with the lymph nodes. Therefore, the likelihood of cancer cells spreading (metastasizing) to other body parts is less likely.

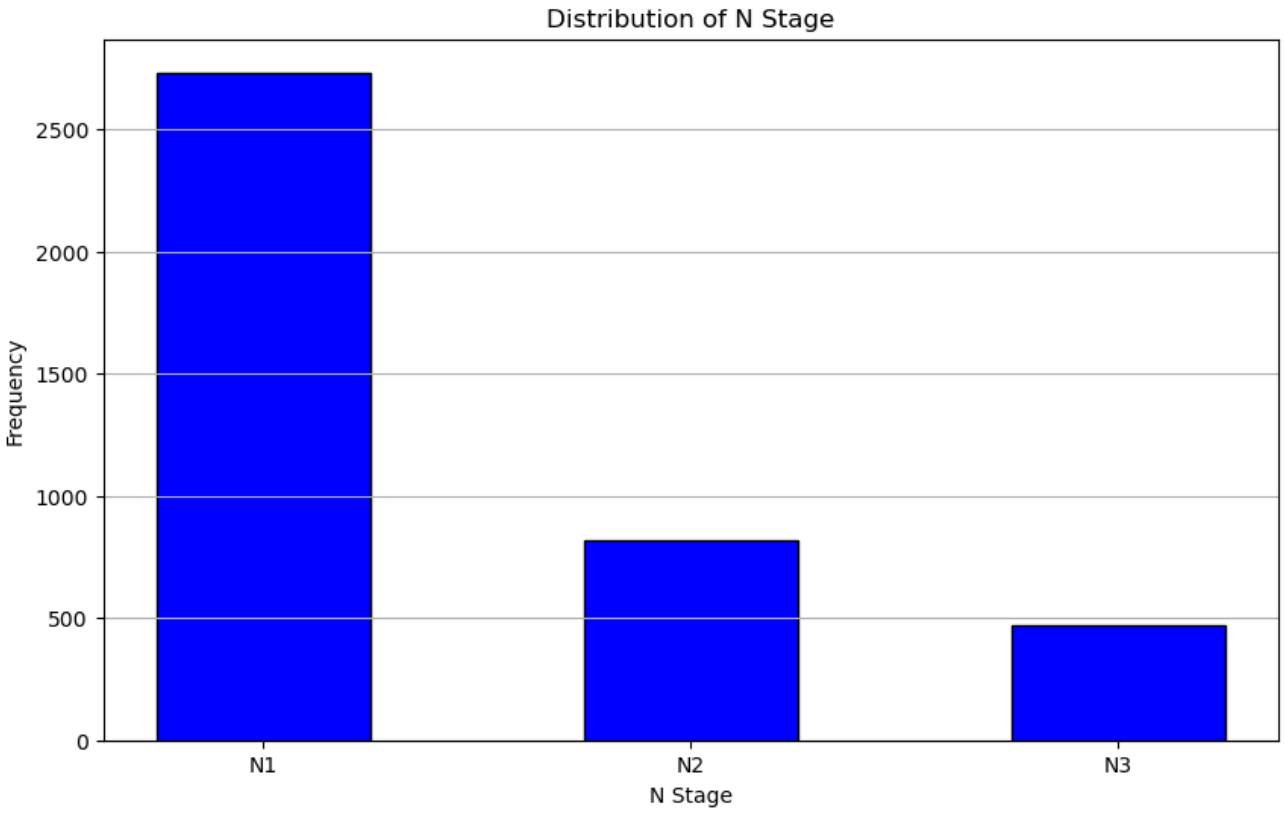
In [761…

```
import pandas as pd
#Create DataFrame
n_stage_counts = pd.DataFrame({
    'Stage': ['N1', 'N2', 'N3'],
    'Count': [2732, 820, 472]
})
#print(n_stage_counts)

#Plotting the distribution of N Stage
plt.figure(figsize=(10,6))
plt.bar(n_stage_counts['Stage'], n_stage_counts['Count'], color='blue', edge
plt.title('Distribution of N Stage')
plt.xlabel('N Stage')
```

```
plt.ylabel('Frequency')
plt.grid(axis='y')

# Show the plot
plt.show()
```



The relationship Between Tumor Size and Survival Months

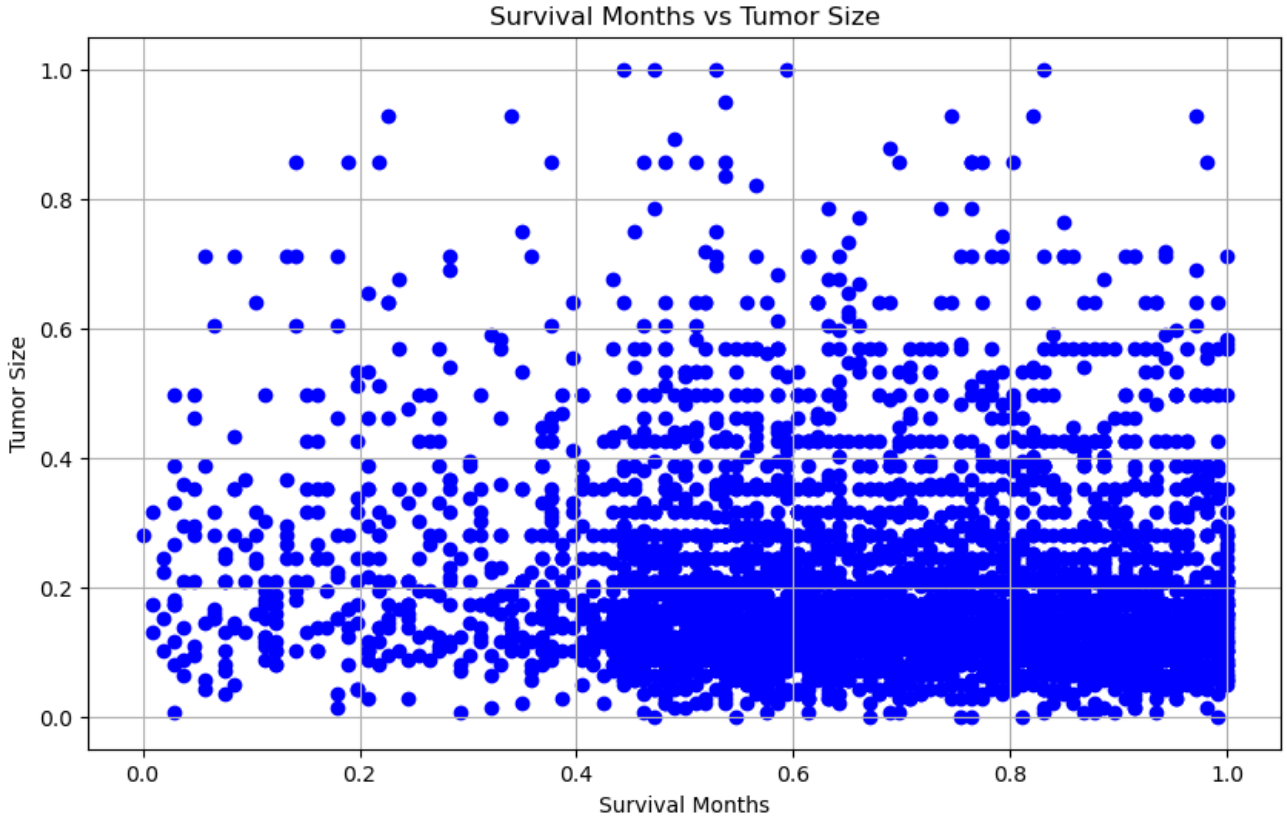
The graph shows that there is somehow an inverse relationship between the tumor size and survival months. It indicates that people with smaller tumor sizes tend to live slightly longer. Breast cancer with a larger tumor size also suggests that the disease is more advanced and often classified as T2 or higher.

In [410…

```
import numpy as np
import pandas as pd
#import matplotlib.pyplot as plt
import seaborn as sns

#Scatter plot of Tumor Size vs Survival Months
plt.figure(figsize=(10, 6))
plt.scatter(x=df_encoded['Survival Months'], y=df_encoded['Tumor Size'], col
plt.title('Survival Months vs Tumor Size')
plt.xlabel('Survival Months')
plt.ylabel('Tumor Size')
```

```
plt.grid()
plt.show()
```



Is there any relationship between regional node-positive and survival months?

The regional node positive indicates the spread of cancer cells to the lymph nodes. The positive lymph node is the indication that cancer cells are being metastasized to other parts of the body, which can be very lethal depending on where and how fast the cancer cells are spreading. The following graph shows that people with breast cancer who have low regional node positives tend to live longer. As the number of nodes with cancer cells increases, the survival months (rate) of people with breast cancer decreases. Therefore, the following graph shows an inverse relationship between Regional Node Positive and Survival Months.

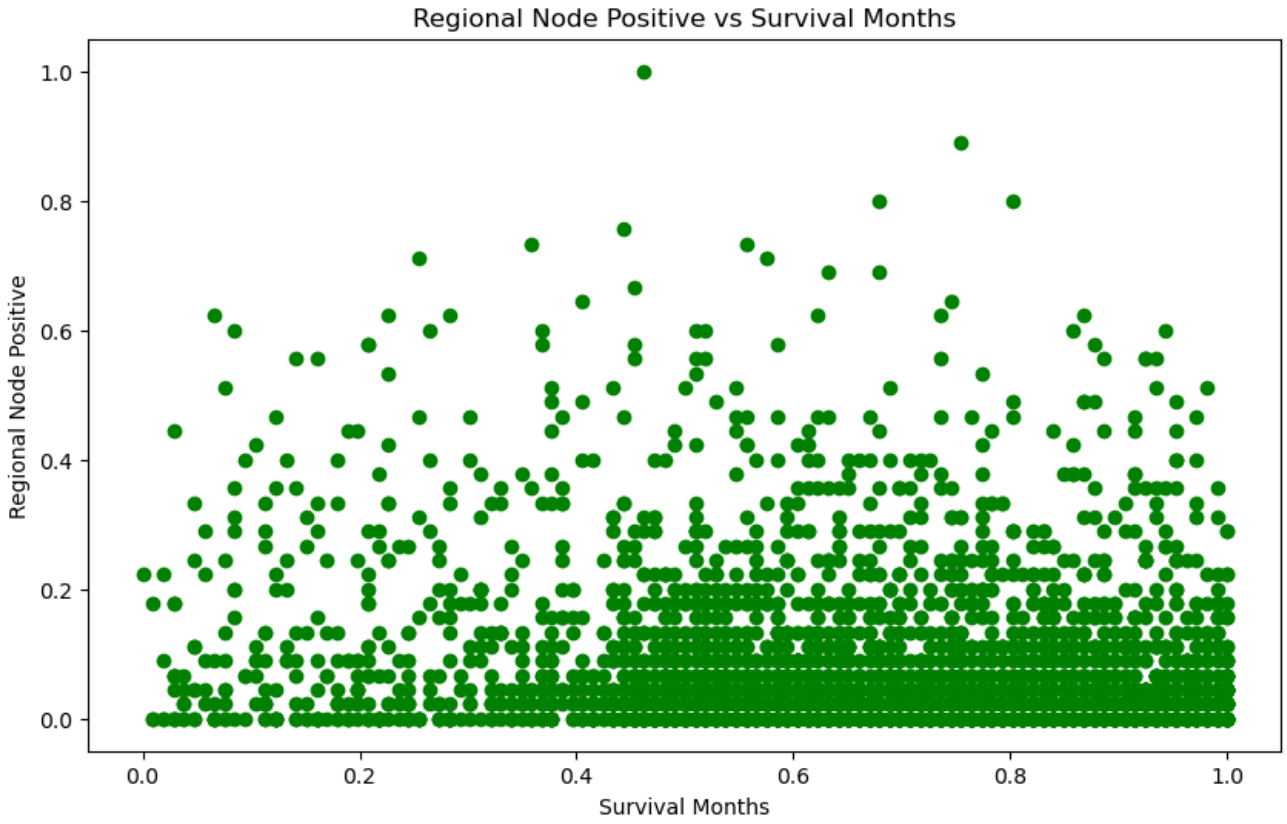
In [ ]:

In [424…

```
import pandas as pd
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
plt.scatter(x=df_encoded['Survival Months'], y = df_encoded['Reginol Node Po
plt.title('Regional Node Positive vs Survival Months')
```

```
plt.xlabel('Survival Months')
plt.ylabel('Regional Node Positive')
plt.show()
```



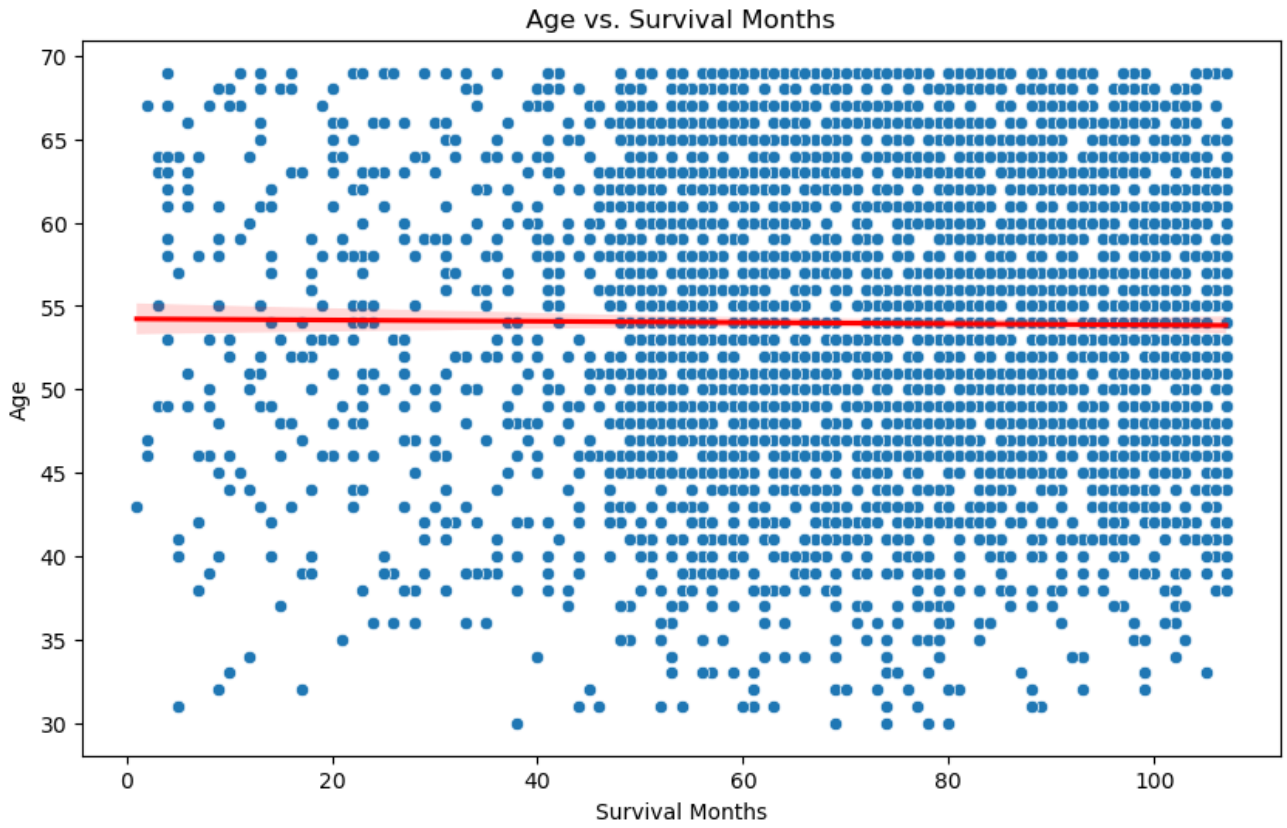
In [ ]:

Age and Survival Months relationships

The graph of age vs. survival months shows little to no correlation and insufficient evidence of a relationship.

In [667…

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Survival Months', y='Age')
sns.regplot(data=df, x='Survival Months', y='Age', color='red', scatter=Fals
plt.title("Age vs. Survival Months")
plt.xlabel("Survival Months")
plt.ylabel("Age")
plt.show()
```



What does the '6th Stage' tell us about the spread of breast cancer?

The "6th Stage" is basically the classification of cancer stages based on the system used to describe cancer spread. These sub-stages described in this study are IIA, IIB, IIIA, IIIB, and IIIC. The first sub-stage IIA shows that the tumor is larger than stage I but did not spread extensively. Even though the cancer is still localized, it might have spread to the nearby lymph nodes. Stage IIIC is characterized by cancer spreading to several lymph nodes (more than 10 lymph nodes), can be large and aggressive, and requires aggressive treatment. These classical sub-stages of cancer help determine the treatment type and predict the outcome.

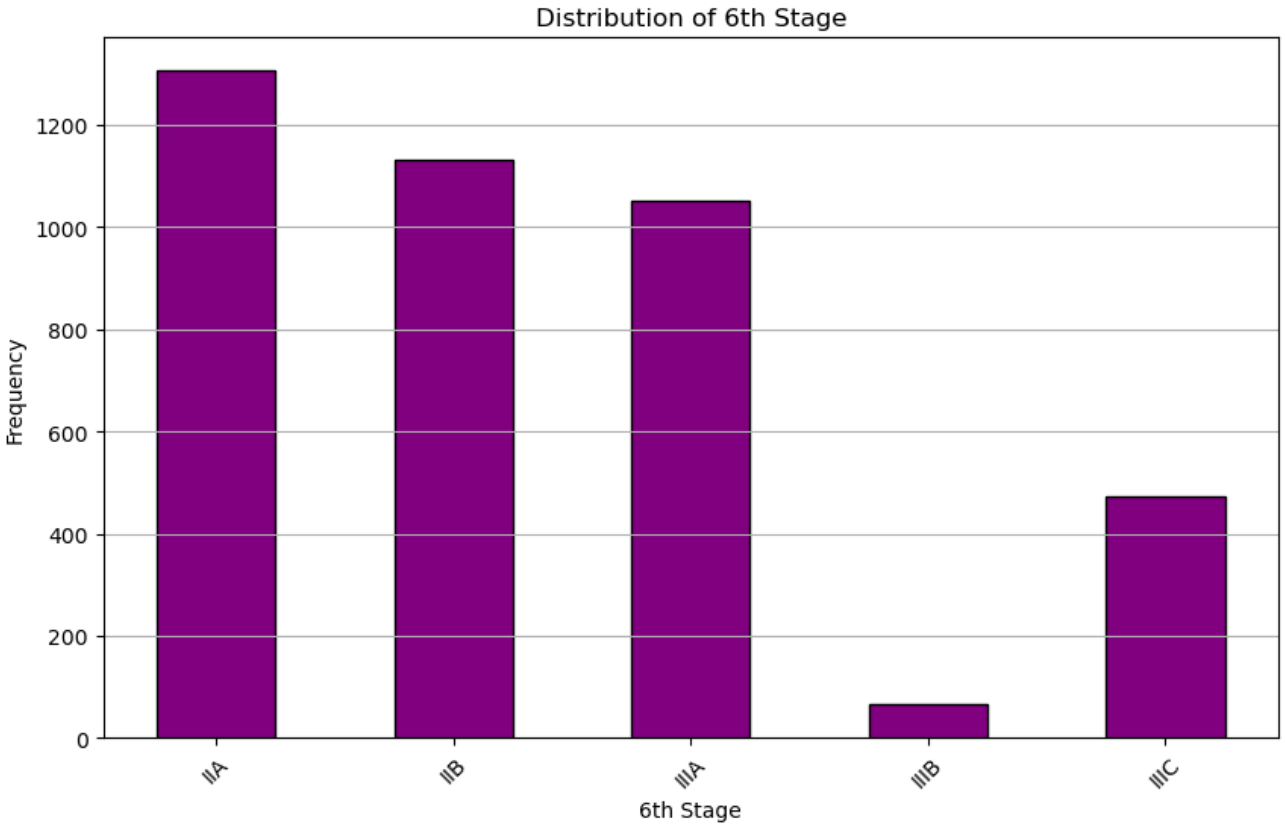
In [645…

```
import pandas as pd
import matplotlib.pyplot as plt
six_stage = pd.DataFrame({
    'Stage':['IIA',  'IIB', 'IIIA', 'IIIB', 'IIIC'],
    'Frequency': [1305, 1130, 1050,   67,  472 ]
})


#bar plot
plt.figure(figsize=(10,6))
plt.bar(six_stage['Stage'], six_stage['Frequency'], color='purple', edgecolo
```

```
plt.title('Distribution of 6th Stage')
plt.xlabel('6th Stage')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.grid(axis='y')

# Show the plot
plt.show()
```



The Status of Estrogen and Progesterone

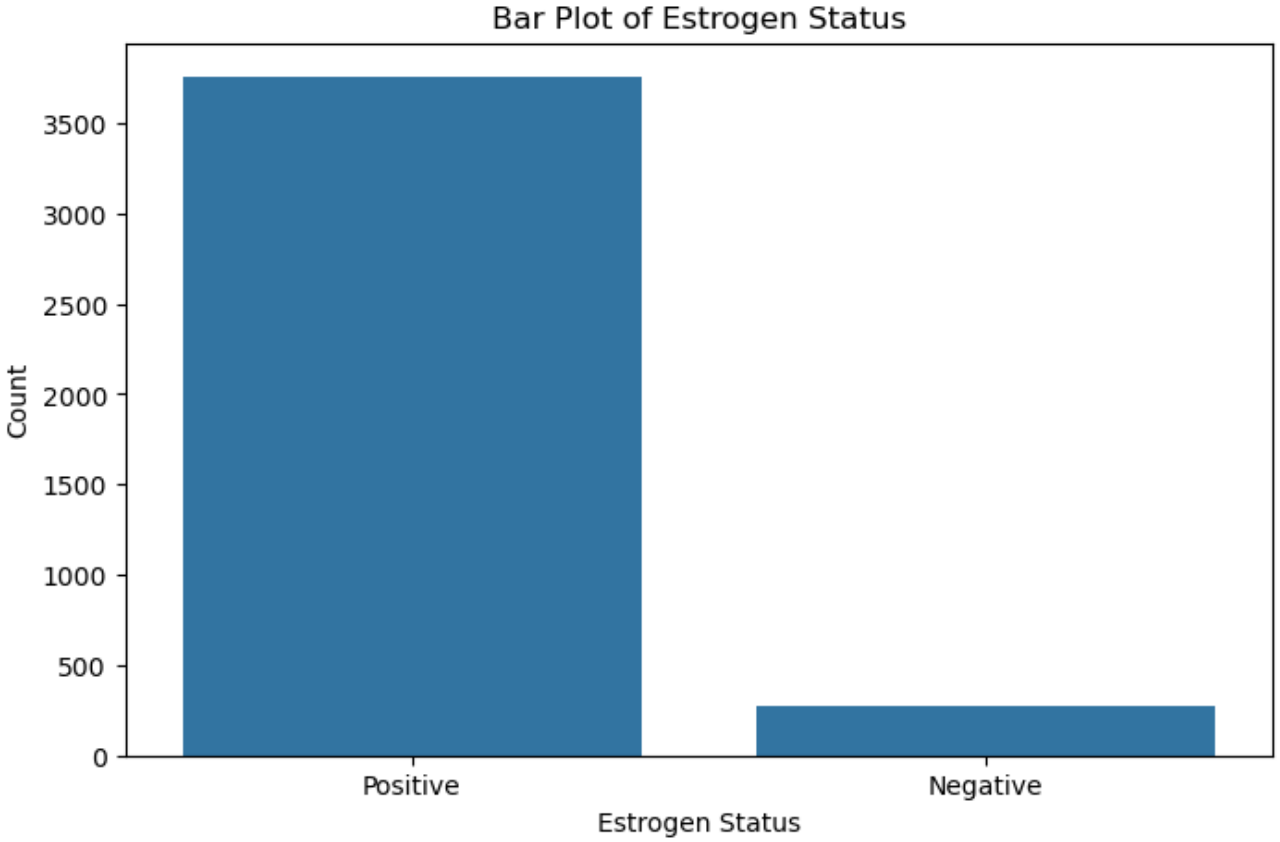
The status of estrogen and progesterone in breast cancer patients plays a significant role in determining the treatment options and predicting the outcome. The majority of this data shows the positive status of estrogen and progesterone, meaning these hormones have receptors that can respond to hormone therapies during treatment. Positive estrogen and progesterone also grow slower than estrogen and progesterone with negative receptors. The negative status of both hormones has fewer treatment options and worse outcomes.

In [763…

```
#Estrogen Status
estrogen = df['Estrogen Status'].value_counts()
frame = pd.DataFrame(estrogen)
```

```
#print(frame)
#plot
plt.figure(figsize=(8,5))
sns.barplot(frame, x = 'Estrogen Status', y = df['Estrogen Status'].value_co
plt.title('Bar Plot of Estrogen Status')
plt.xlabel('Estrogen Status')
plt.ylabel('Count')
plt.show()
```

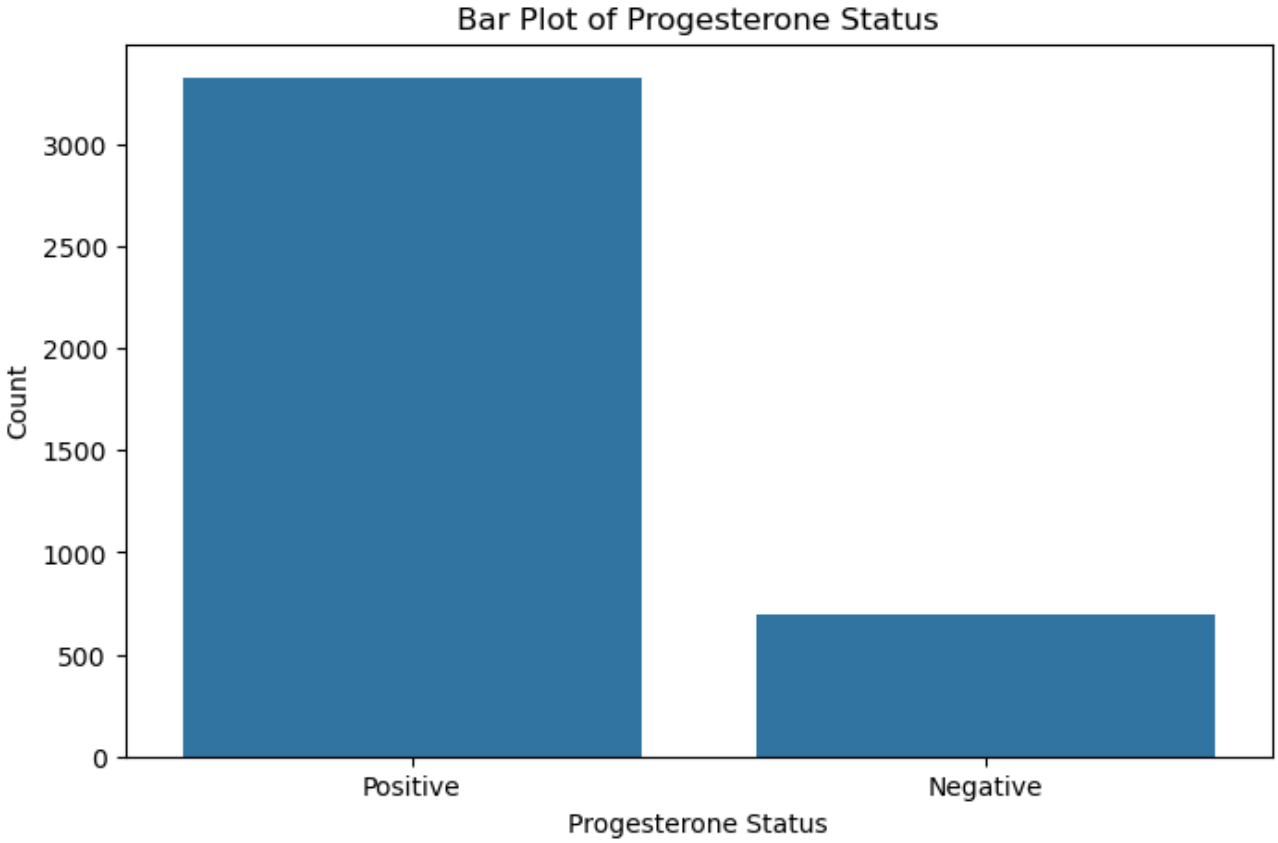
```
                 count
Estrogen Status
Positive          3755
Negative           269
```



In [765…

```
#Progesterone Status
progesterone = df['Progesterone Status'].value_counts()
frame1 = pd.DataFrame(progesterone)
print(frame1)
plt.figure(figsize=(8,5))
sns.barplot(frame1, x = 'Progesterone Status', y = df['Progesterone Status']
plt.title('Bar Plot of Progesterone Status')
plt.xlabel('Progesterone Status')
plt.ylabel('Count')
plt.show()
```

```
                     count
Progesterone Status
Positive              3326
Negative               698
```



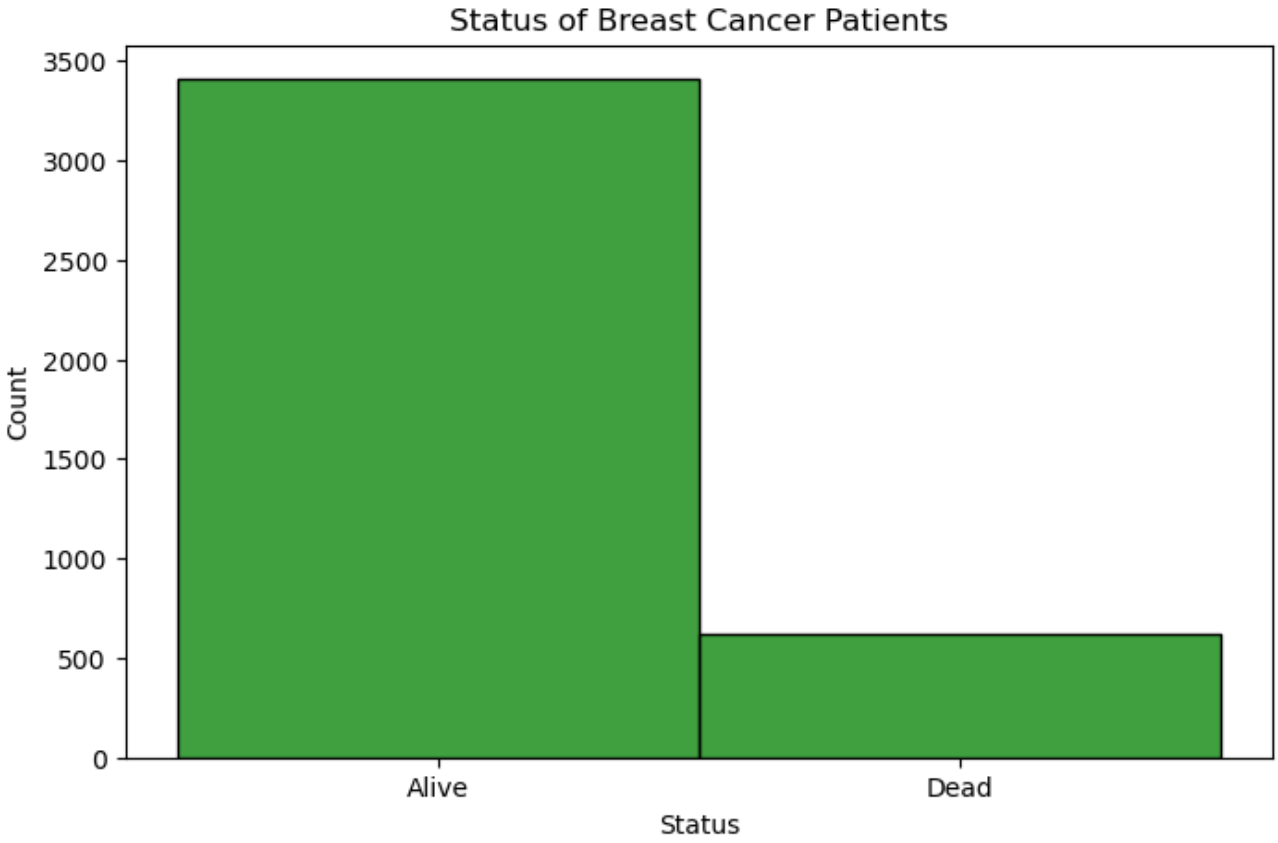
Status

In [767…

```
#Status
status = df.Status.value_counts()
frame =pd.DataFrame(status)
#print(frame) #prints the data frame

#bar plot
plt.figure(figsize=(8,5))
sns.histplot(df['Status'], color = 'green')
plt.title('Status of Breast Cancer Patients')
plt.xlabel('Status')
plt.ylabel('Count')
plt.show()
```

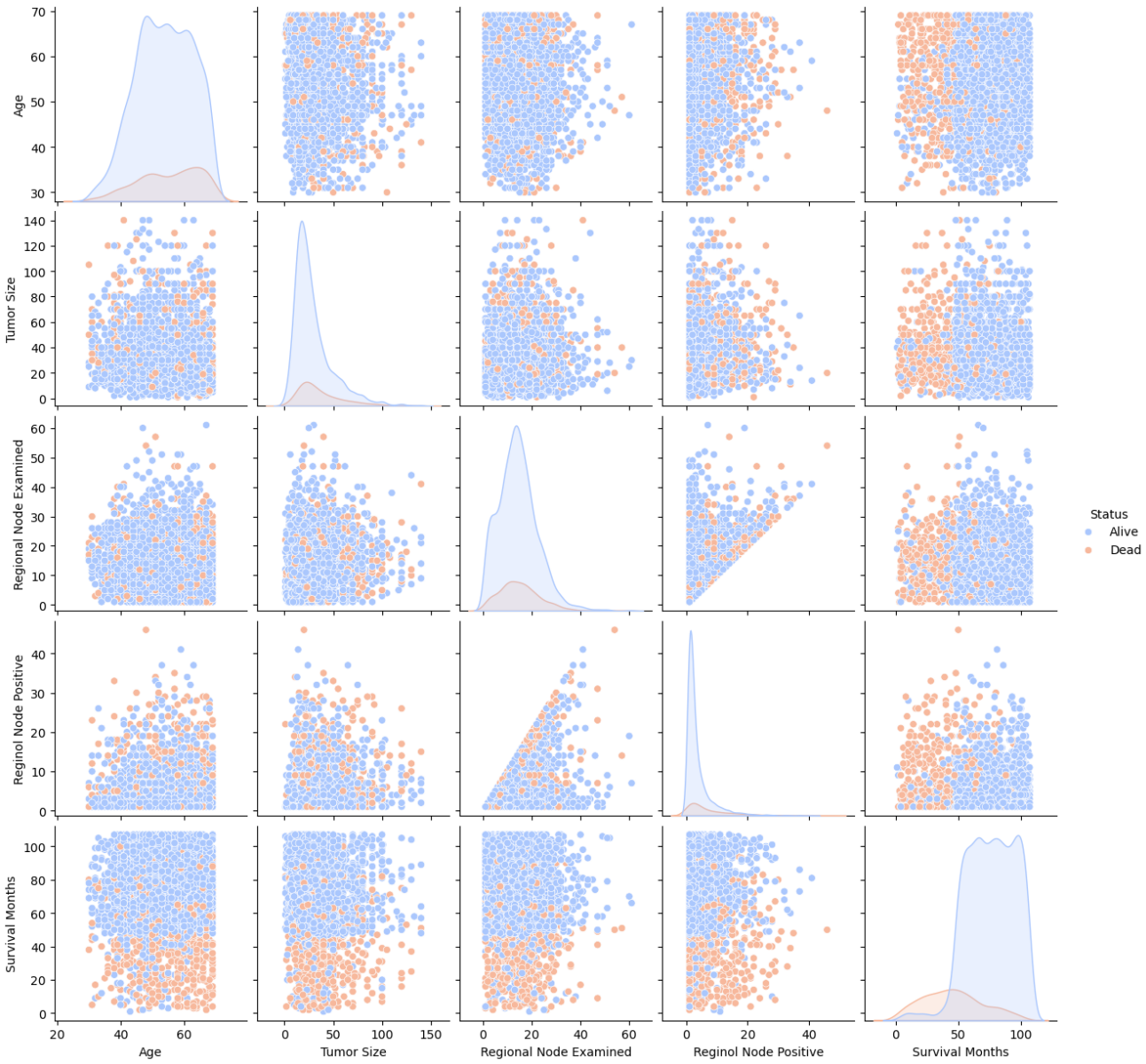
```
        count
Status
Alive    3408
Dead      616
```



Conclusion

In conclusion, I want to use 'sns.pairplot()' to visualize the overall relationship of all variables based on the 'Status' of the patients.

In [719…
```
#plt.figure(figsize=(10,6))
sns.pairplot(df, hue ="Status", palette ='coolwarm')
plt.show()
```



References

In [521…

```
# APA citation for the website
citation = (
    "Namdari, R. (Nov 2017). Breast cancer dataset. Kaggle. "
    "https://www.kaggle.com/datasets/reihanenamdari/breast-cancer"
)
citation1 = (
    "National Cancer Institute. (n.d.). Breast cancer treatment (PDQ®)–Healt
    "U.S. Department of Health and Human Services. https://www.cancer.gov/ty
)

print(citation)
print(citation1)
```

Namdari, R. (Nov 2017). Breast cancer dataset. Kaggle. https://www.kaggle.com/datasets/reihanenamdari/breast-cancer
National Cancer Institute. (n.d.). Breast cancer treatment (PDQ®)–Health professional version. U.S. Department of Health and Human Services. https://www.cancer.gov/types/breast/patient/breast-treatment-pdq#_148