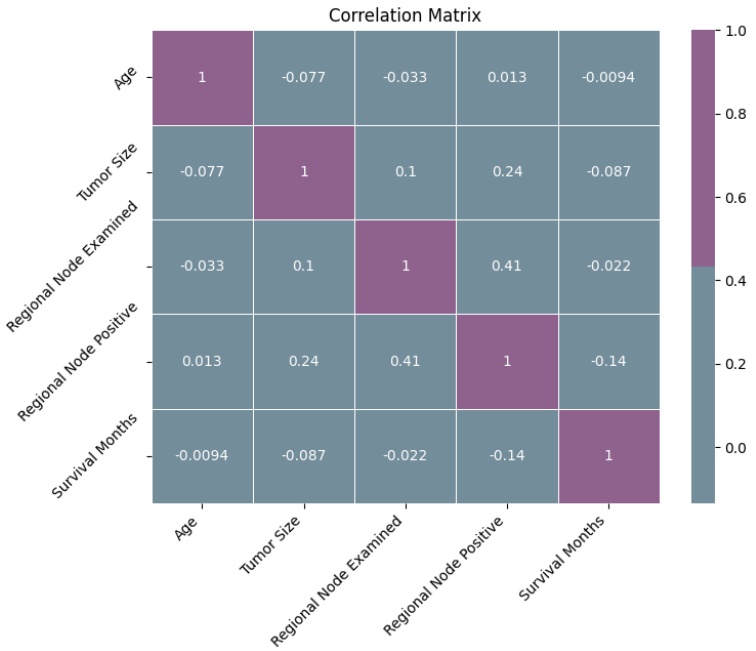
Appendix

A.1: Correlation Matrix - Raw Dataset

As another step in the preprocessing part, we compute the correlation matrix to see how variables interact with each other. Next you can see the final result:



We see no strong correlation between numerical variables, some insights we can get, for example, are the following:

- Considering the correlation between *Age* and *Tumor Size*, we can see a weak correlation of -0.077. This indicates that as the age increases, the tumor size tends to decrease slightly, but, again, this correlation is really weak
- The correlation between *Regional Node Examined* and *Regional Node Positive* indicates a moderate positive correlation of 0.412, thus, as the number of regional nodes examined increases, the number of positive nodes also tends to increase
- The correlation between *Survival Months* and other variables is relatively weak. This indicates a weak or no linear relationship between survival months and the other numerical variables in the dataset.

A.2: Outlier Analysis Results' for Numerical Variables

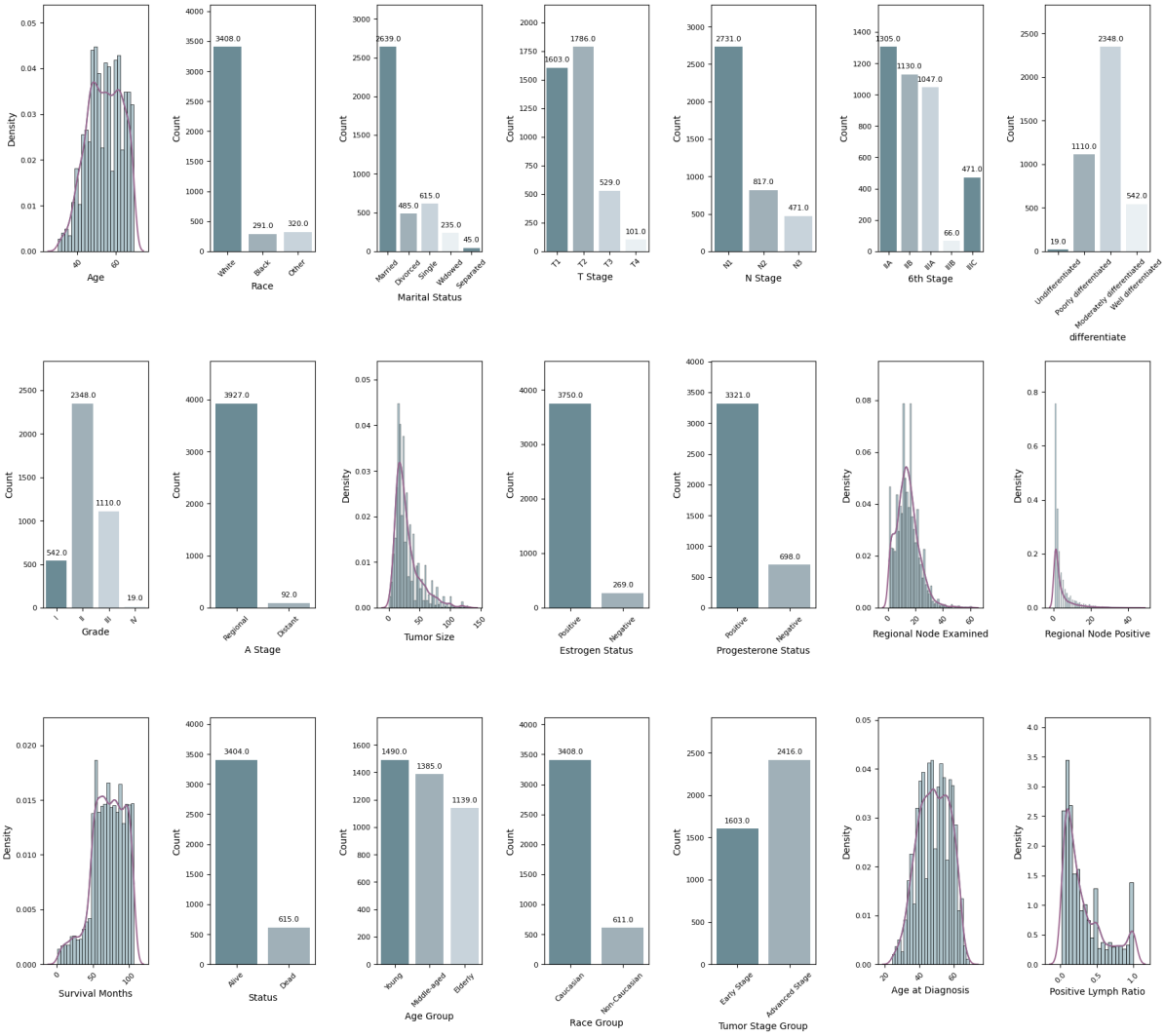
```
Tumor Size Outlier Analysis:
----------------------------
First Quartile (Q1): 16.00
Third Quartile (Q3): 38.00
Interquartile Range (IQR): 22.00
Lower Bound: -17.00
Upper Bound: 71.00
Outliers Length: 222
Extreme Outliers Length: 5
```

```
Regional Node Examined Outlier Analysis:
----------------------------
First Quartile (Q1): 9.00
Third Quartile (Q3): 19.00
Interquartile Range (IQR): 10.00
Lower Bound: -6.00
Upper Bound: 34.00
Outliers Length: 72
Extreme Outliers Length: 0
```

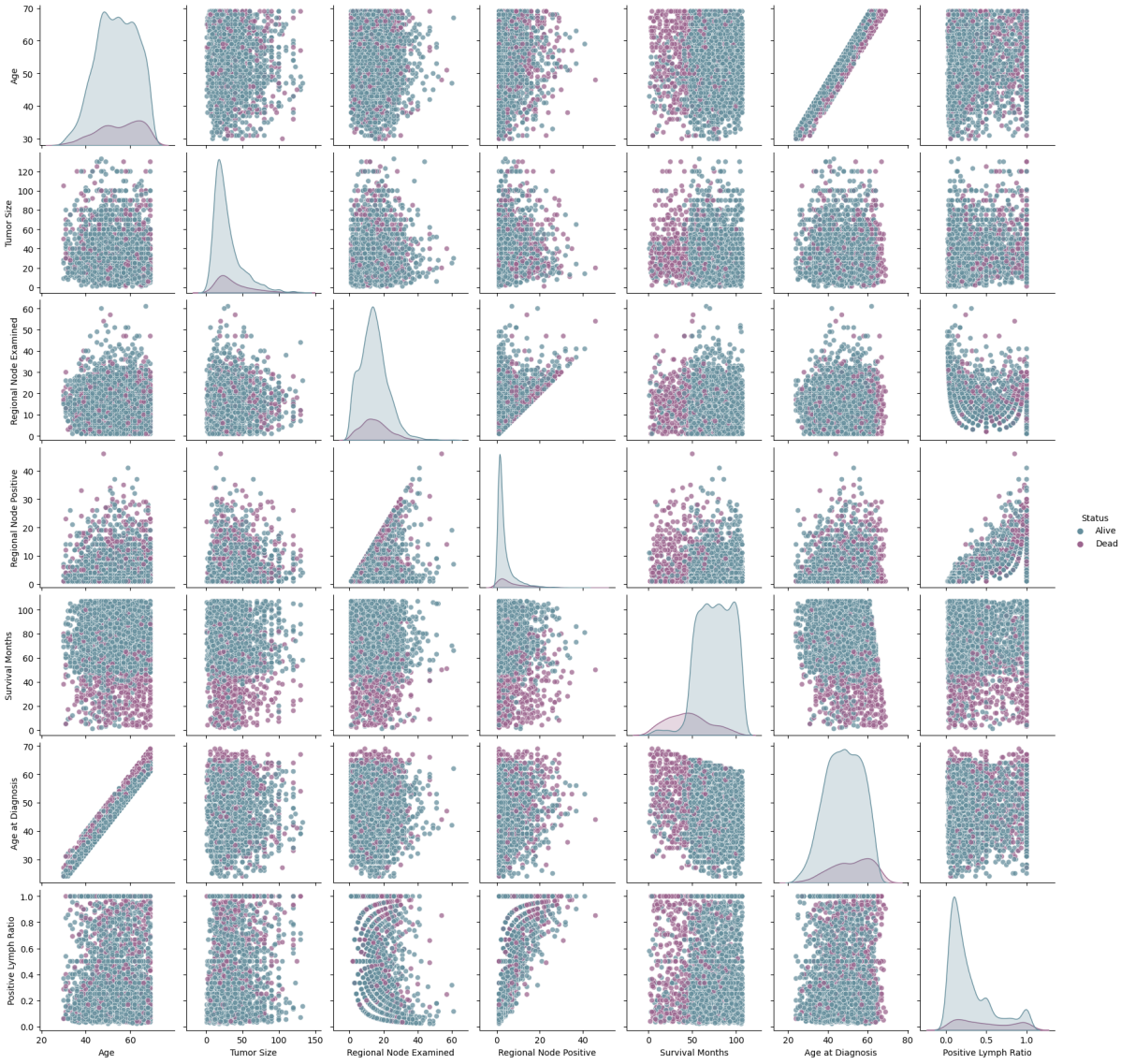
```
Regional Node Positive Outlier Analysis:
----------------------------
First Quartile (Q1): 1.00
Third Quartile (Q3): 5.00
Interquartile Range (IQR): 4.00
Lower Bound: -5.00
Upper Bound: 11.00
Outliers Length: 344
Extreme Outliers Length: 54
```

```
Survival Months Outlier Analysis:
----------------------------
First Quartile (Q1): 56.00
Third Quartile (Q3): 90.00
Interquartile Range (IQR): 34.00
Lower Bound: 5.00
Upper Bound: 141.00
Outliers Length: 18
Extreme Outliers Length: 0
```

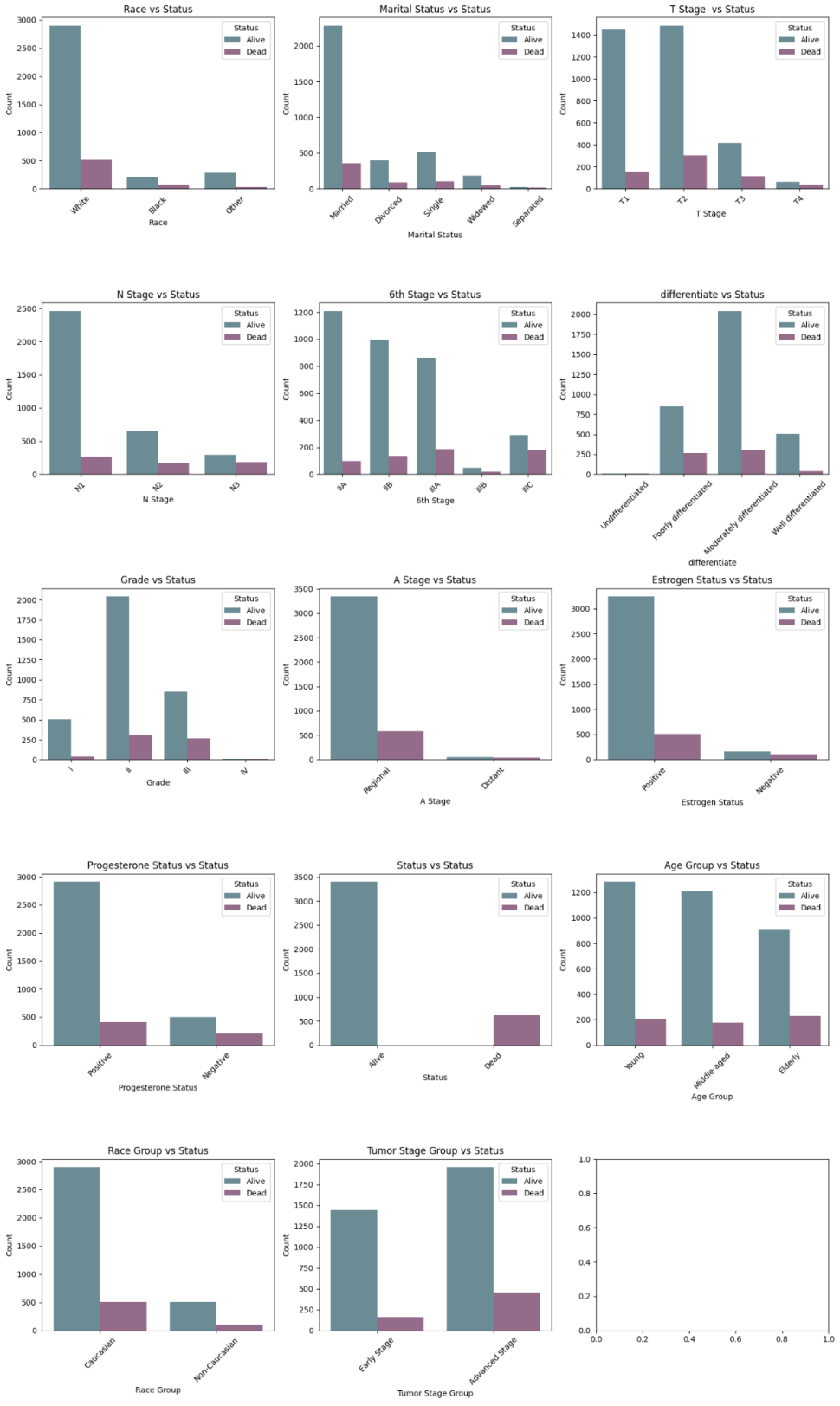
A.3: Univariate Analysis - Final Dataset



A.4: Bivariate Scatterplot Analysis with Respect to Status - Final Dataset



A.5: Bivariate Barplot Analysis with Respect to Status - Final Dataset



A.6: Correlation Matrix - Final Dataset