Seminar 3

Yufeng Deng Yuanqing Wang

Task1

1.1 Feature Selection & Model Development

First, we calculated the correlation coefficient between each variable. From figure 1.1 we can see that the correlation values between TC-LDL, HDL-APOA1, WBC-NEU are almost 1, which means these pairs are highly linearly related, leading to multicollinearity. We used Elastic Net to simplify the model and reduce the probability of multicollinearity.

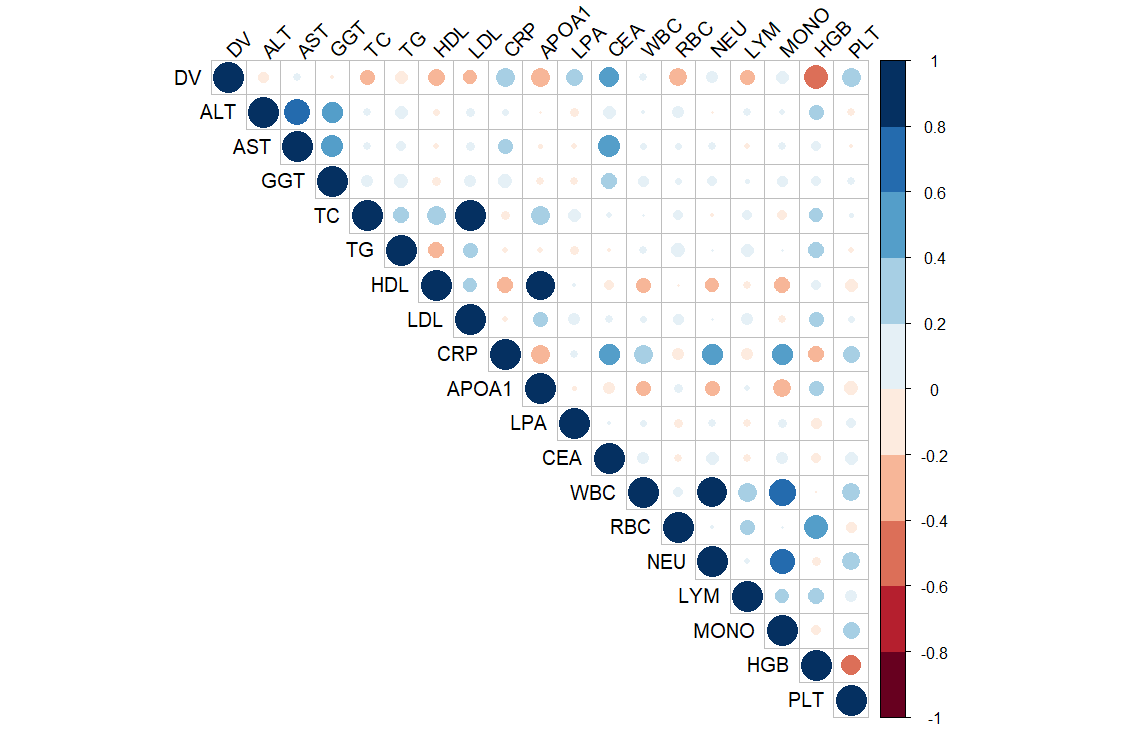


Figure 1.1 Correlation Matrix

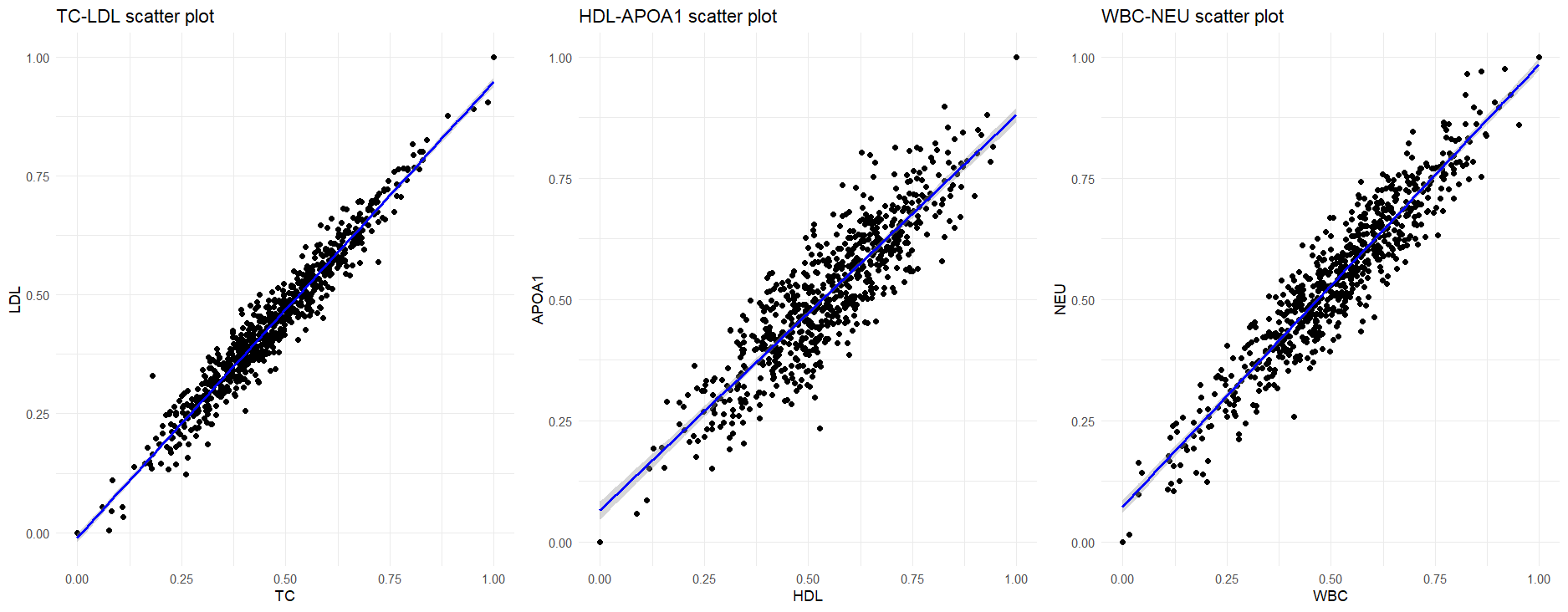


Figure 1.2 Linear Relationship

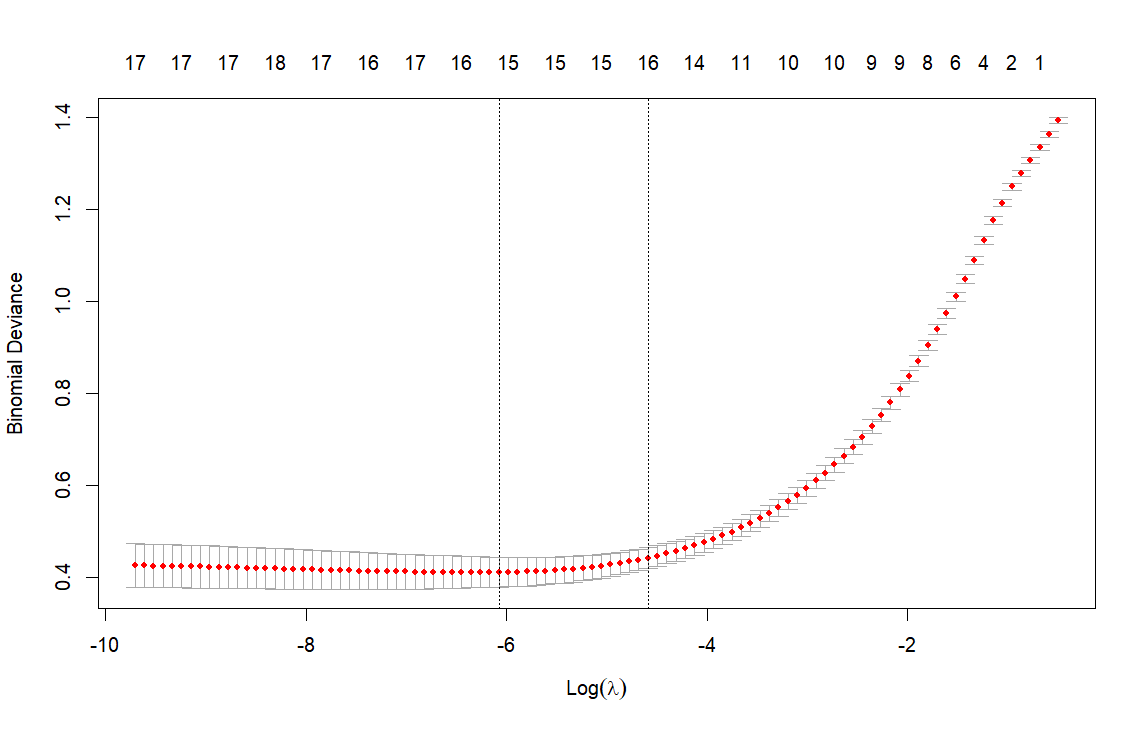


Figure 1.3 Cross Validation of Elastic Net

We used the package ‘glmnet’ to implement Elastic Net. Figure 1.3 shows how does it find the best lambda value, which is 0.002294098 in this case. After Regression, the coefficients are shown in Table 1.1. It obvious that the regression deleted the features GGT, TC and CRP, and gave other features weight coefficients. Using this result, we constructed a linear model to do the binary classification task.

Table 1.1 Coefficients of Elastic Net

|  |  |
| --- | --- |
| Feature | Coefficient |
| Intercept | 0.1096291 |
| ALT | -0.2006522 |
| AST | -0.3814531 |
| GGT | 0 |
| TC | 0 |
| TG | -0.1771271 |
| HDL | -1.2138212 |
| LDL | -0.8198785 |
| CRP | 0 |
| APOA1 | -0.1409414 |
| LPA | 1.6898303 |
| CEA | 2.5336365 |
| WBC | -0.1975546 |
| RBC | 0.1212922 |
| NEU | -0.2468349 |
| LYM | -0.8486259 |
| MONO | 0.2116491 |
| HGB | -2.5728811 |
| PLT | 0.7236508 |

1.2 Diagnose Model Performance.

We calculated several Evaluation Metrics of the prediction of our model and compared them with the prediction of the linear model without feature selection. Table 1.2 shows the results, and we can see that with Elastic Net, all the Metrics of the linear model increased.

Table 1.2 Evaluation Metrics of the prediction (transform threshold is 0.5)

|  |  |  |
| --- | --- | --- |
| Metrics | Before Elastic Net | After Elastic Net |
| Accuracy | 0.8832 | 0.9112 |
| Precision | 0.8558 | 0.8846 |
| Recall | 0.8990 | 0.9293 |
| F1 | 0.8768 | 0.9064 |

1.3 Threshold Value

In diagnosis, detecting a CRC patient as a healthy person (False Negative) is more costly than detecting a healthy person as a CRC patient (False Positive). So, thresholds may need to be biased to increase the Recall rates to reduce the likelihood of False Negative. To achieve that, we can lower the threshold of the transform fromprobability to binary class. For instance, when we turn the threshold from 0.5 (default) to 0.3, the Recall rate of our model will increase from 0.9293 to 0.9596.

1.4 Clinical Implementation

In clinical implementation, doctors can check the coefficients of the model like the Table 1.1, and they can decide to manually delete a feature by considering other information, and they also can modify the value based on their knowledge, because this model is simple and easy to explain. On the other hand, if we know a certain patient’s information, we can select the data from patients with similar physical conditions to increase the performance of the model. This may involve privacy risk, so the hospitals need to consider of it and develop a reasonable system to protect the data.

Task 2

1. Analyze trends in CRC stage

We checked the dataset, the number of people in each stage is equal, all are 50. The distributions of Age are different among all the stages. From Table 2.1 and Figure 2.1, we can briefly tell that the trend is people detected in later stage are older. That means when getting older, the probability of late-stage detection increases. For others, the values are shown in Table 2.2 ~ 2.4 and Figure 2.2 ~ 2.4. We applied Chi-square tests for these categorical variables. And we concluded that All the three variables have statistically significant association with the stage.

Table 2.1 Age distribution by Stage

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | Mean Age | SD Age | Min Age | Max Age |
| 1 | 26.1 | 3.78 | 21 | 37 |
| 2 | 29.4 | 3.56 | 22 | 37 |
| 3 | 31.9 | 3.77 | 23 | 39 |
| 4 | 36.0 | 3.78 | 28 | 43 |

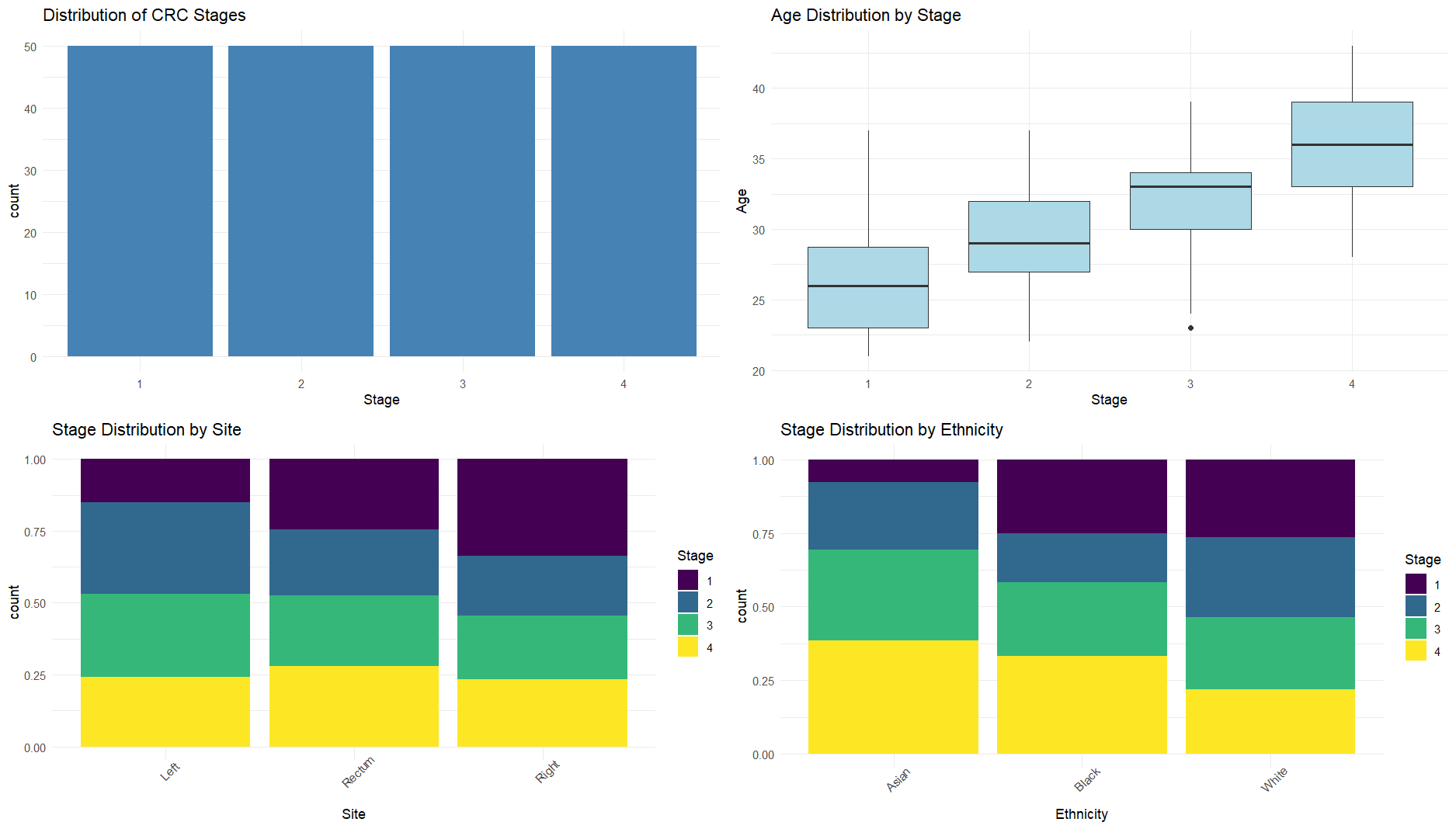


Figure 2.1 Age distribution by Stage

Regarding racial distribution, we found that the sample primarily consists of White individuals. We noted that ethnic minorities (Black and Asian populations) appear to have a higher proportion of late-stage diagnoses. Concerning the impact of lifestyle, we discovered that individuals living alone are more likely to receive late-stage diagnoses compared to those with partners. We believe this may be related to Mutual care between partners. In terms of tumor location characteristics, we observed that rectal cancer cases show a relatively high proportion of late-stage diagnoses. We found that right-sided colon cancers have a slightly better early detection rate compared to left-sided colon cancers.

Table 2.2 Stage distribution by Lifestyle

|  |  |  |
| --- | --- | --- |
| Stage | Alone | Partnered |
| 1 | 20 | 30 |
| 2 | 21 | 29 |
| 3 | 27 | 23 |
| 4 | 29 | 21 |

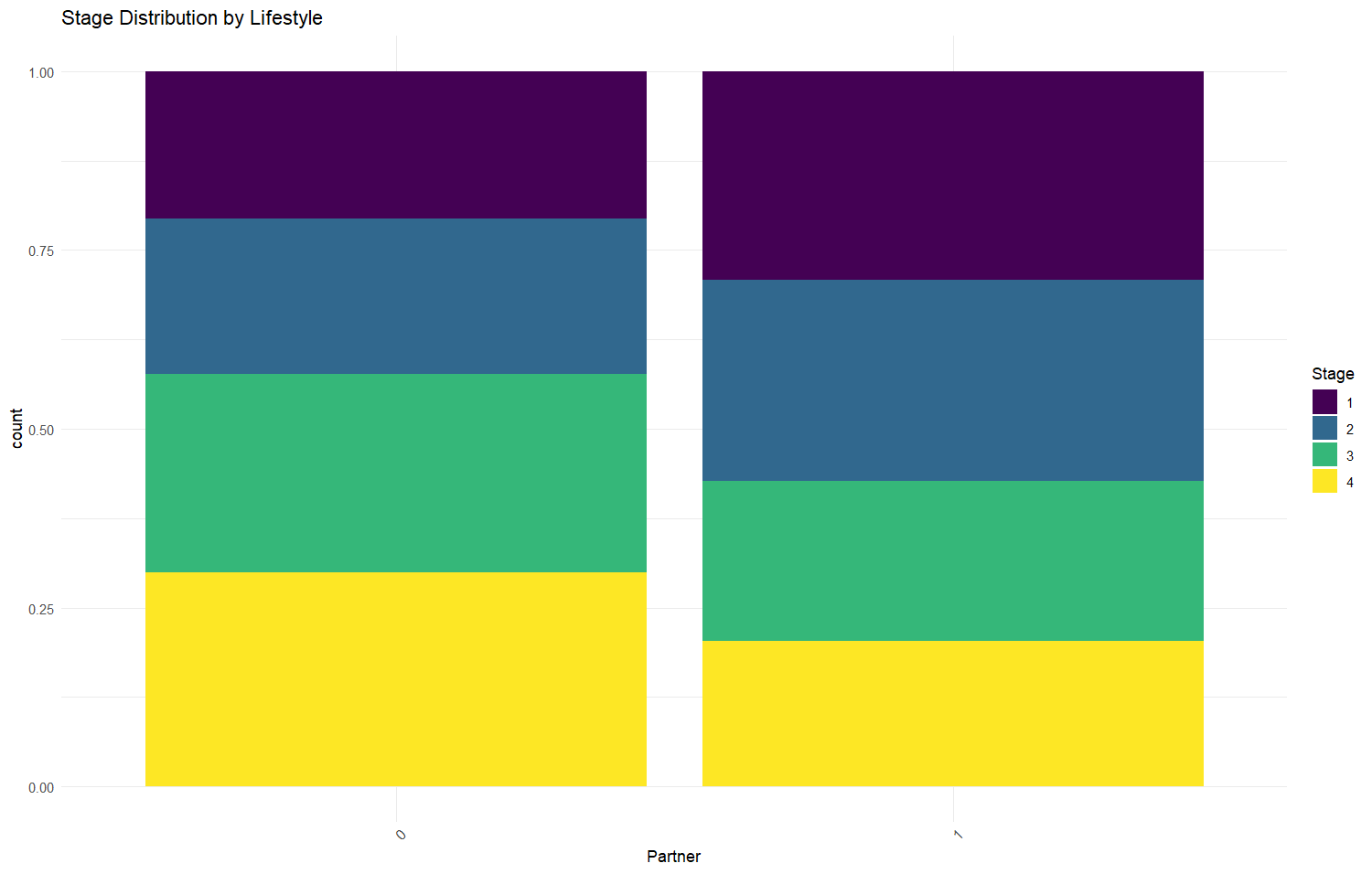


Figure 2.2 Stage distribution by Lifestyle

Table 2.3 Stage distribution by Ethnicity

|  |  |  |  |
| --- | --- | --- | --- |
| Stage | Asian | Black | White |
| 1 | 1 | 9 | 40 |
| 2 | 3 | 6 | 41 |
| 3 | 4 | 9 | 37 |
| 4 | 5 | 12 | 33 |

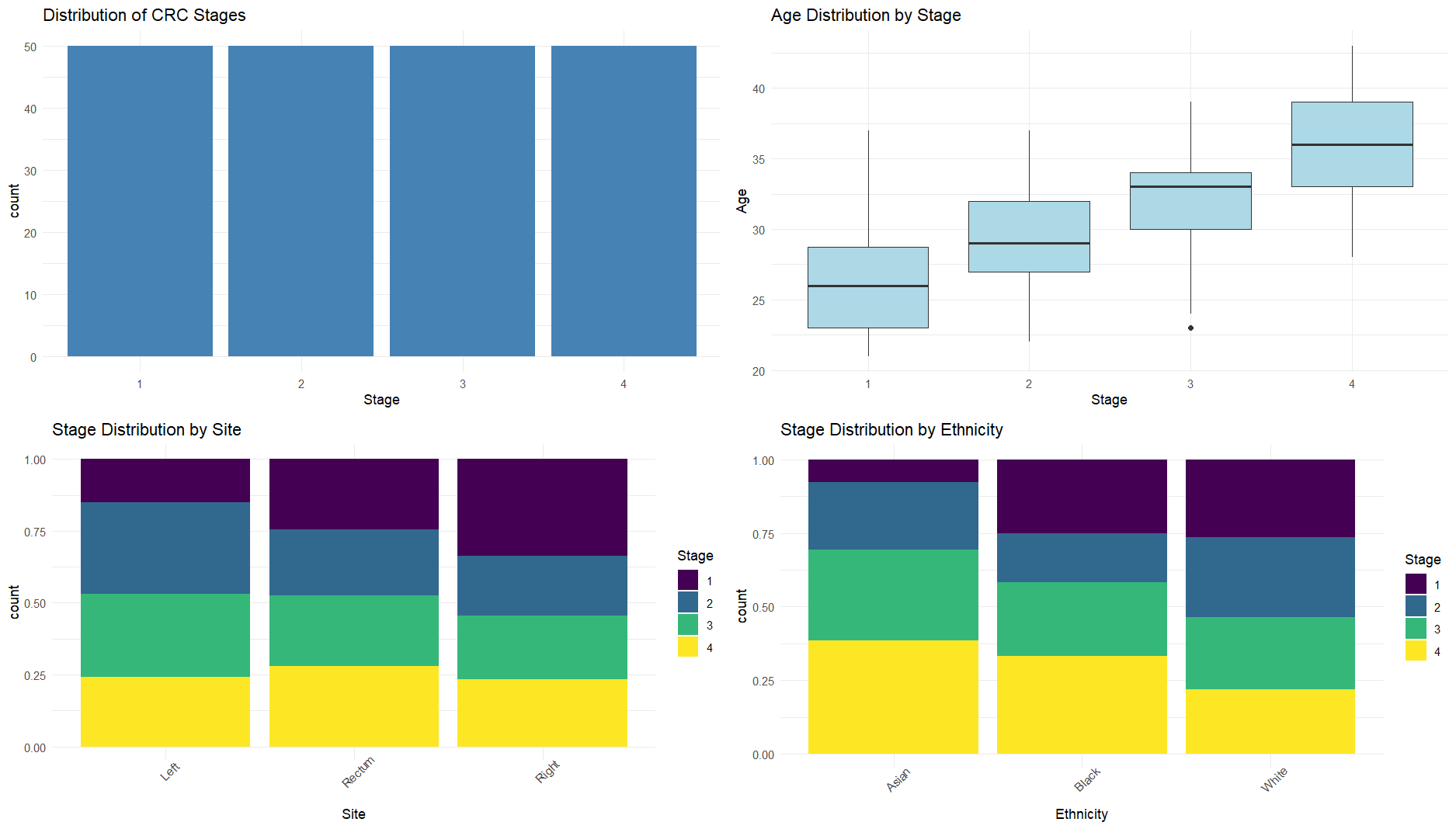


Figure 2.3 Stage distribution by Ethnicity

Table 2.4 Stage distribution by Site

|  |  |  |  |
| --- | --- | --- | --- |
| Stage | Left | Rectum | Right |
| 1 | 10 | 14 | 26 |
| 2 | 21 | 13 | 16 |
| 3 | 19 | 14 | 17 |
| 4 | 16 | 16 | 18 |

图表, 条形图

描述已自动生成

Figure 2.4 Stage distribution by Site

2. Recommendations based on the data analysis

According to the Age impact, we recommend strengthening early screening efforts for CRC and lowering the screening starting age to 20-25 years old. Besides, we strongly recommend to regularly visit young people for diagnosis, especially for those living alone. In terms of public health strategies, we advocate developing targeted health education for ethnic minorities and considering racial differences in healthcare resource allocation. For researchers, we suggest conducting in-depth research on disease mechanisms in young populations and different mechanisms among different cancer site.