SEER Project-Survival time analysis(2 years)

Group 11: Chi Zhang, Aoyi Li, Simu Huang, Zixuan Liu

2021/5/5

Introduction

Over the recent years, there has been an explosion in the use of machine learning to be applied in our real life, especially in the public domain and business to predict behaviors and investment decisions. In the clinical area, machine learning can also be used to predict the survival time of the cancer patients, which will greatly benefit early diagnosis, clinical management of cancer patients, and treatment adjustment.

In this project, we mainly focus on the factors that affect the survival time of the head and neck cancer patients after diagnosis and use machine learning methods to predict whether head and neck cancer patients will survive longer than two years after being diagnosed. Besides, what we notice was that in the data, the average survival time of the patients is greater than two years. We will also try to identify the factors that affect the survival of patients less than two years after diagnosis and do some descriptive research by EDA to find the relationship between these variables.

Data Processing

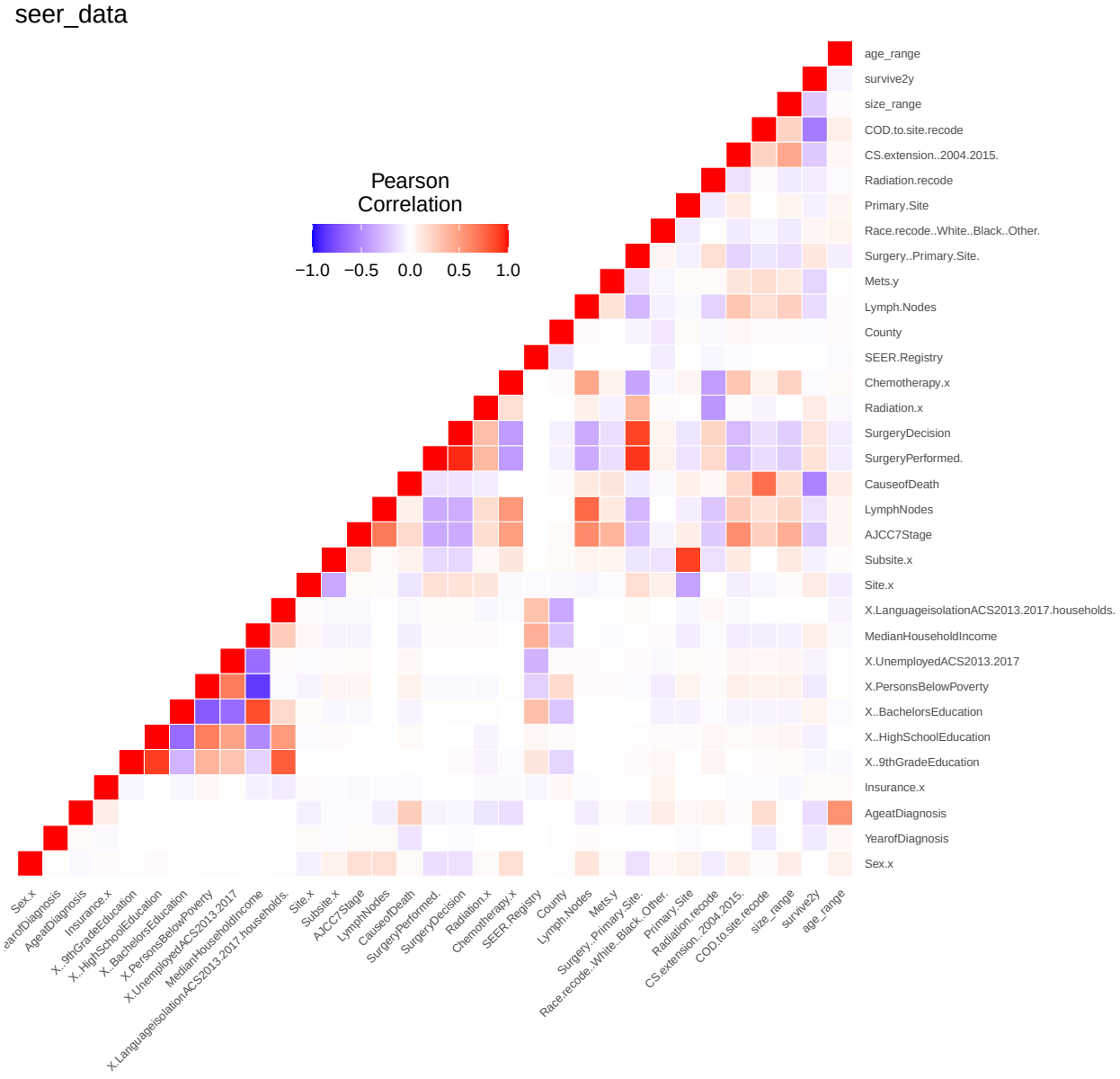
The Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. The raw data of all head and neck cancer has many missing values, but it has the column survival_month, our target variable. We also have two other files containing the information which raw data does not have. To make sure we can use all necessary information at once, we merged raw data with two other datasets by study ID. What is more, we extracted the data only from 2010 to 2014, since the data before 2010 has many blanks and the data after 2014 is not sufficient for our 2 years survival analysis.

To better perform our models, we transferred all factor variables into numeric. We also split the data into two parts that 70% as training data and 30% as testing data. We will train our models on training data and test the validation on testing data later on.

EDA

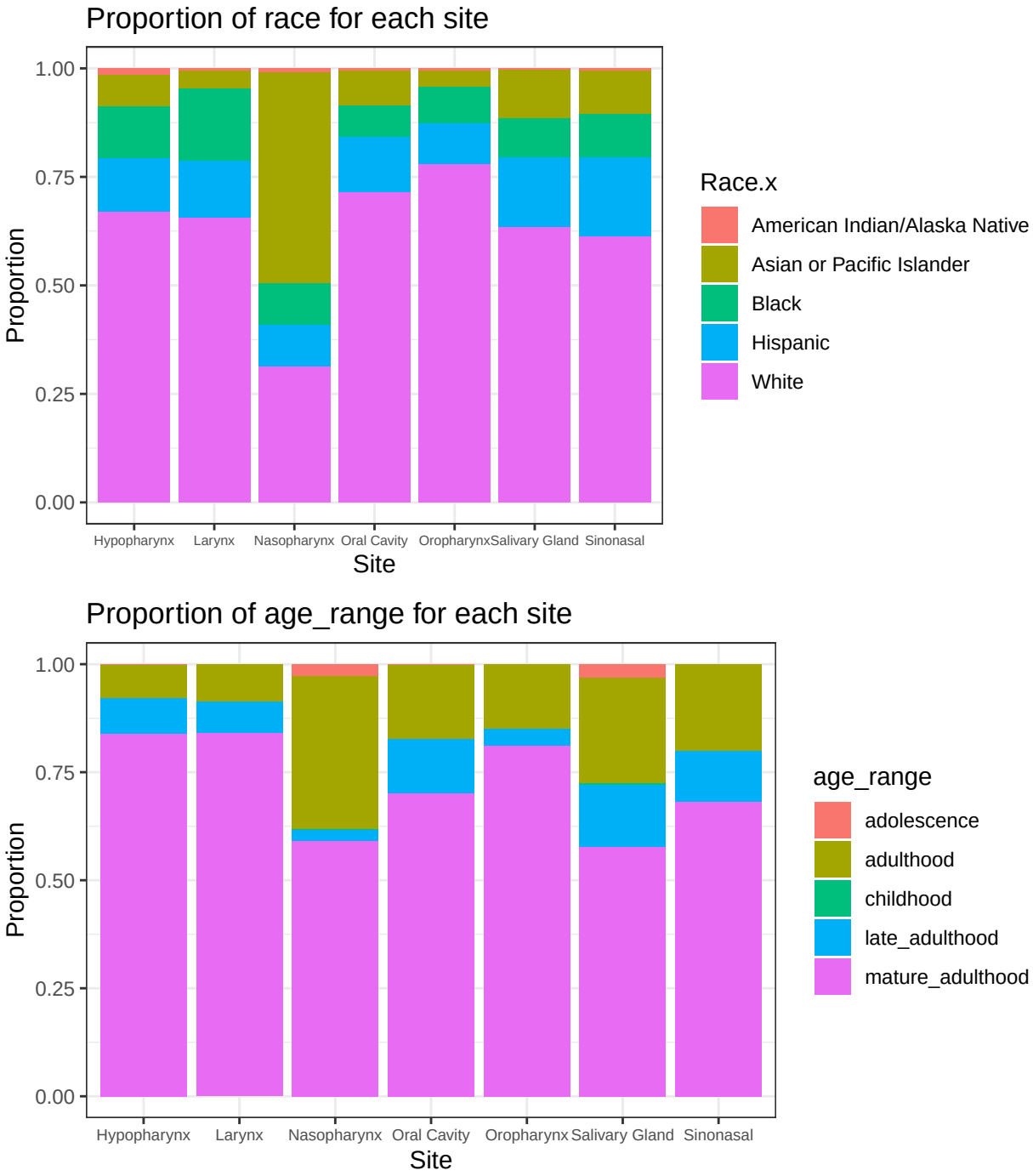
Heatmap Plot

Shown in below is a correlation map for the seer data that describes the relationship between the different features.



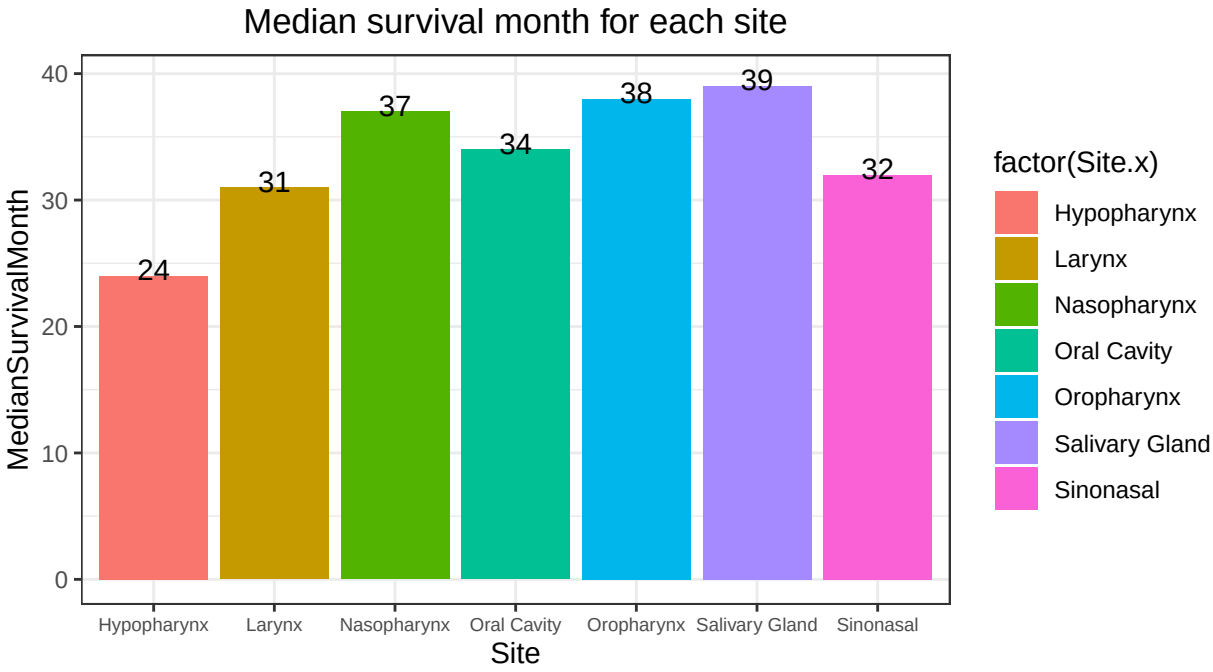
From the correlation heatmap, we can see that there is some high correlation between variables. There could exist collinearity which may affect modeling. From the plot that the variable person below poverty and median household income have high negative correlation. That makes sense since people with lower incomes are poorer. We also noticed that few variables have high positive correlation. For example, the variable surgery decision and surgery primary site, which can explain the correlation that if as the bigger the tumor, the more likely people are to undergo surgery.

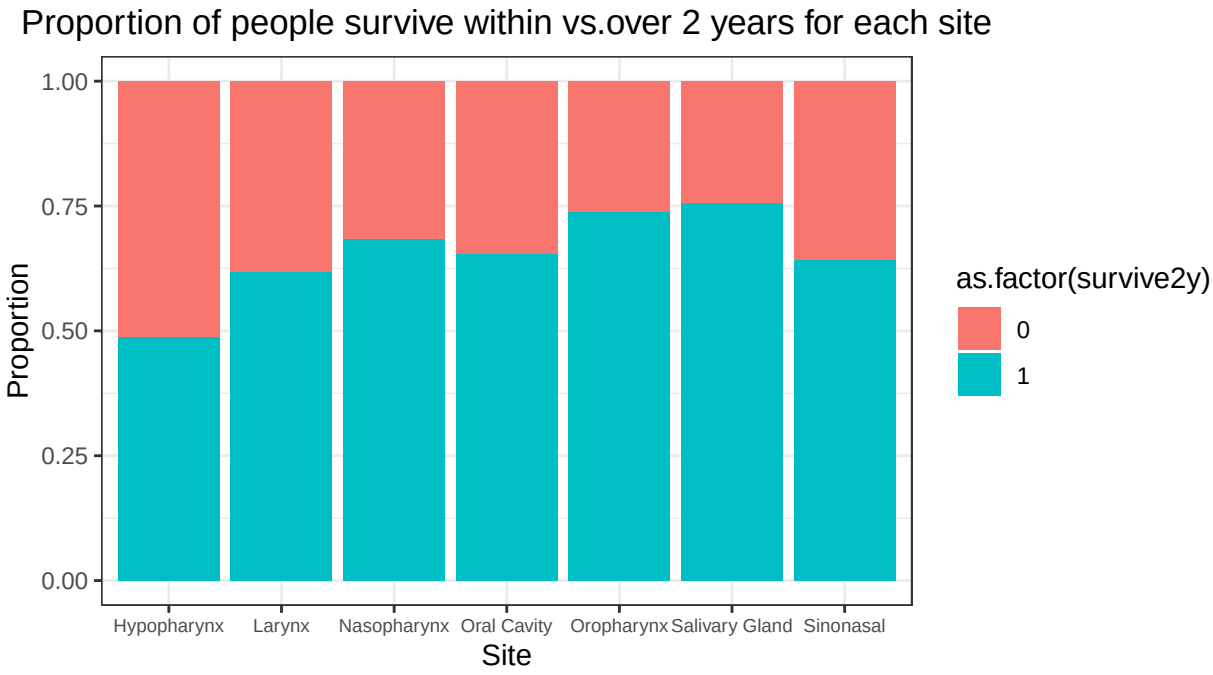
Interesting Findings



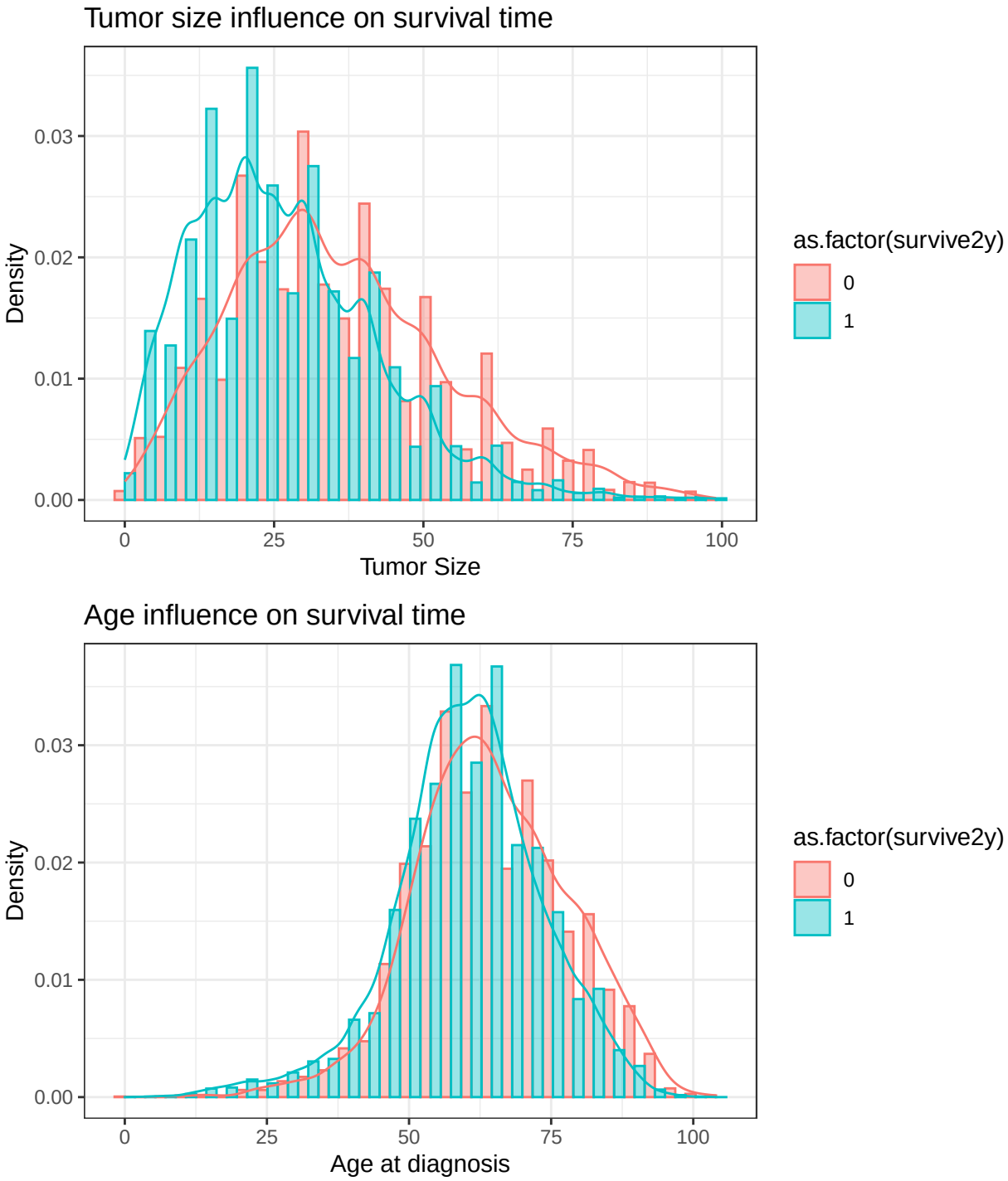
For the 2 plots above, we found that the site called Nasopharynx is obviously different from other sites. From the plot at the top, it is clear that Asian or Pacific Islander people are more likely to have Nasopharynx cancer than other races, and white people are less likely to have this cancer compared to other races. From the plot at the bottom, we can see that adults (age from 21-50) are more likely to have Nasopharynx cancer than other age stages. Then we found out that this is not the exceptional case happened in our dataset. There are some researches show that Asian or Pacific Islander people are more likely to have Nasopharynx cancer than other races.

Survive within/over 2 years

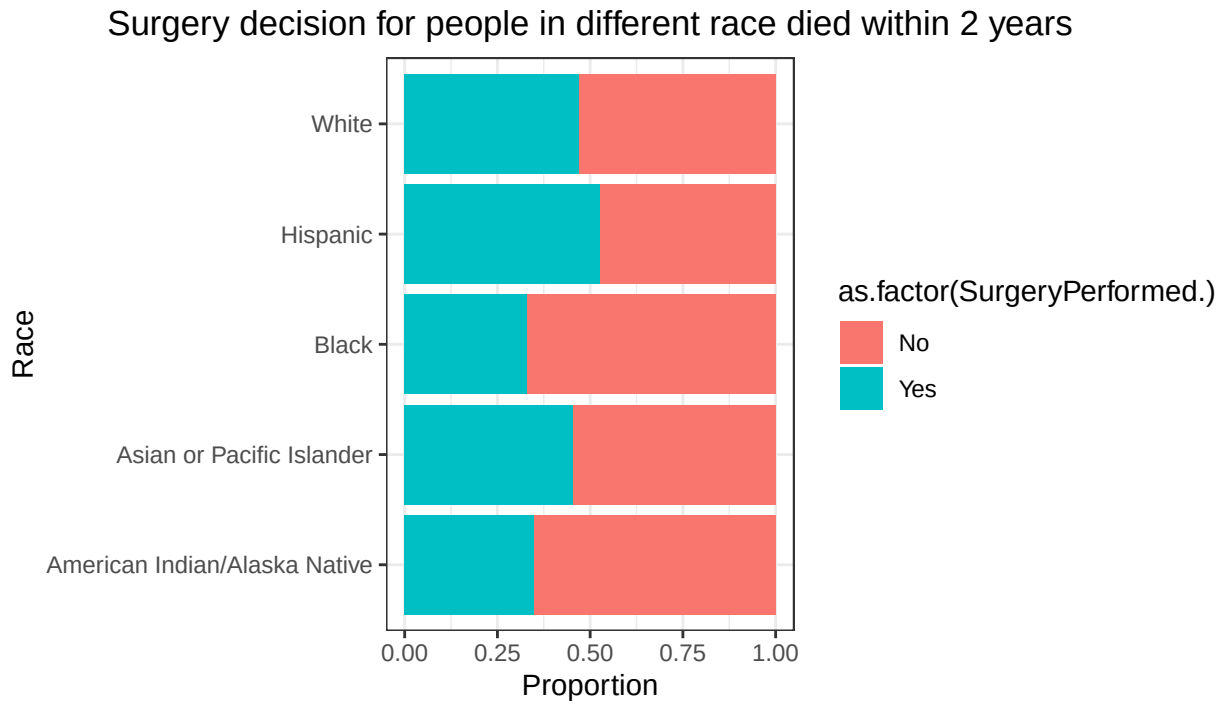


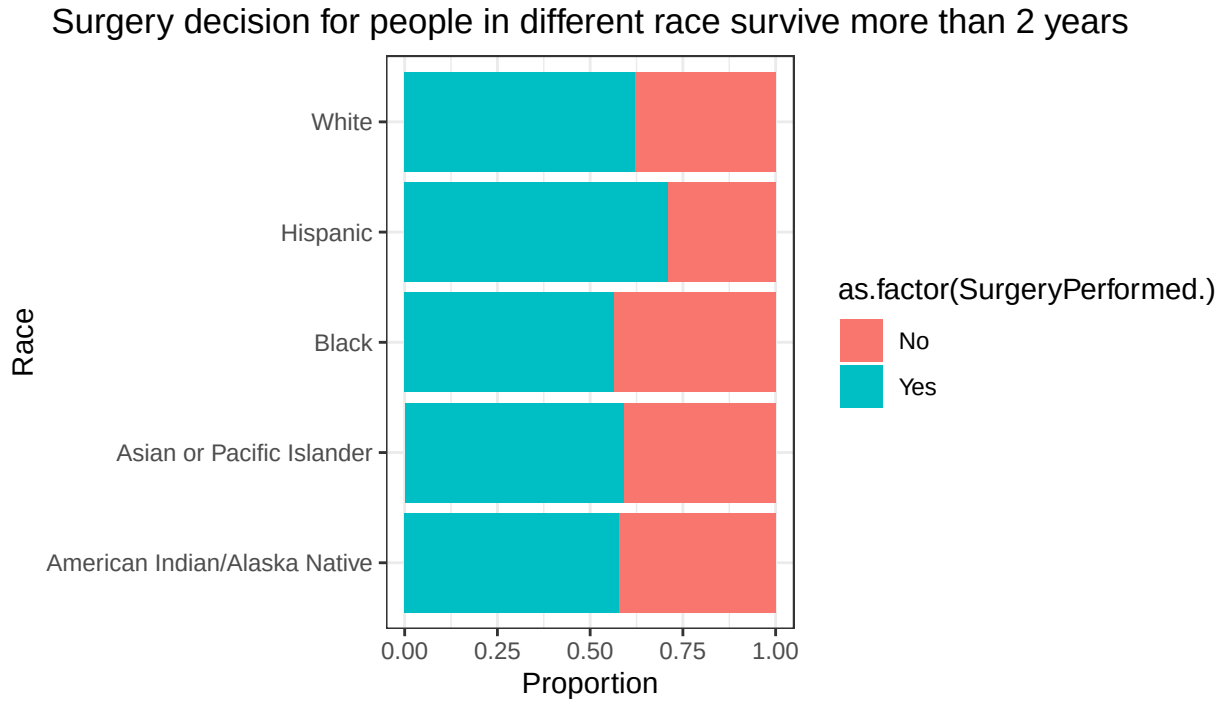


We plotted the median survival months for each site, which shows that almost all cancers here have median survival months more than 2 years. Thus, we would like to predict whether a person can survive over 2 years or not after diagnosis. And further understand that if there is any similarity of people who died within 2 years. Both 2 plots above show that people who have Hypopharynx cancer have the lowest survival time and they are most likely to die within 2 years.



These 2 density plots show influences of tumor size and age on survival time. When tumor sizes are less than 30, people are more likely to survive more than 2 years. However, when sizes are greater than 30, people are more likely to die within 2 years. For the age plot, the differences between ages are not that clear. People who are in or older than the mature adulthood stage have a larger chance to die within 2 years compared to younger people.





Then we would like to check if surgery performance is a significant factor that affects survival time. By comparing these 2 plots, we can clearly see the influence of surgery. It is clear that more than half of people who died within 2 years did not have surgery. For people who survive more than 2 years, a large proportion of them chose to do surgery. Thus, we can conclude that surgery performance should be an important factor.

Model Selection and Validation

Logistic Regression

We first tried with logistic regression to do the prediction. Logistic regression is a generalized linear model. It can be used to classify a binary dependent variable. In our case, it's to classify 0 and 1 which represent patients who have survived less than 24 month or more than 24 month. We did model selection before running into the final model for the logistic regression. We first conducted a model containing all the variables and looked into the summary, took out a few variables that had high p-value and refit the model. For example, we took out education level since we don't think education level has an effect on our predictor and it also has high p-value. After model selection, we improved 0.3511% of our accuracy.

Our final model is written as:

```
## make the model with Logistic
lg<-glm(survive2y~.,family=binomial(link='logit'),data=trainm)
```

XGBoost

```
# one-hot coding
adummies <- as.data.frame(model.matrix(~.-1, afactor)) %>%
select(-217)
combined <- cbind(adummies,anumeric)
# split test and train
numberOfTrainingSamples <- round(nrow(combined) * .7)
train_data <- combined[1:numberOfTrainingSamples,]
test_data <- combined[-(1:numberOfTrainingSamples),]
label_train <- data$survive2y[1:numberOfTrainingSamples]
label_test <- data$survive2y[-(1:numberOfTrainingSamples)]
# put our testing & training data into two seperates Dmatrixs objects
dtrain <- xgb.DMatrix(data = as.matrix(train_data), label= label_train)
dtest <- xgb.DMatrix(data = as.matrix(test_data))
```

We also try to use XGBoost to predict whether patients can survive more than two years after diagnosis. XGBoost(Extreme Gradient Boosting) is a tree-based integrated machine learning algorithm especially where speed and accuracy are concerned. The reason we consider using this algorithm for prediction is that there are a large number of observations and many classification features in the data. After cleaning the data, we convert the categorical variables into numeric using one hot encoding and use xgb.DMatrix to convert the data table into a matrix. XGBoost model needs parameter adjustment to improve its performance. We use the default booster type "gbtree" to help us solve the classification problem. Meanwhile, when the value of "nround" is greater than 200, the accuracy of this model is not significantly improved. Thus, we set "nround" equals 200 in this function, which means that the algorithm will generate 200 decision trees in the final model.

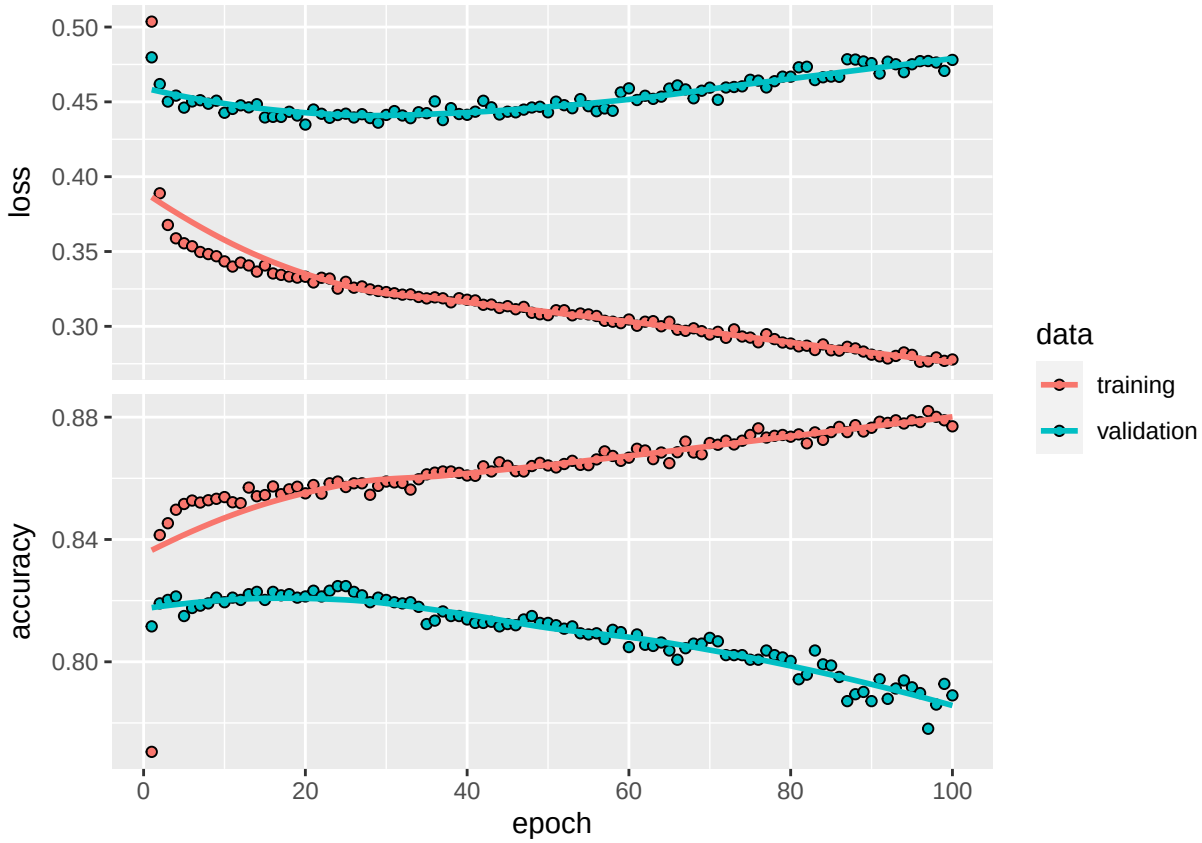
The function is written as:

```
#Xgboost Model
model_x <- xgboost(data = dtrain, # the data
nround = 200, # max number of boosting iterations
objective = "binary:hinge")
```

Multi-layer Perceptron

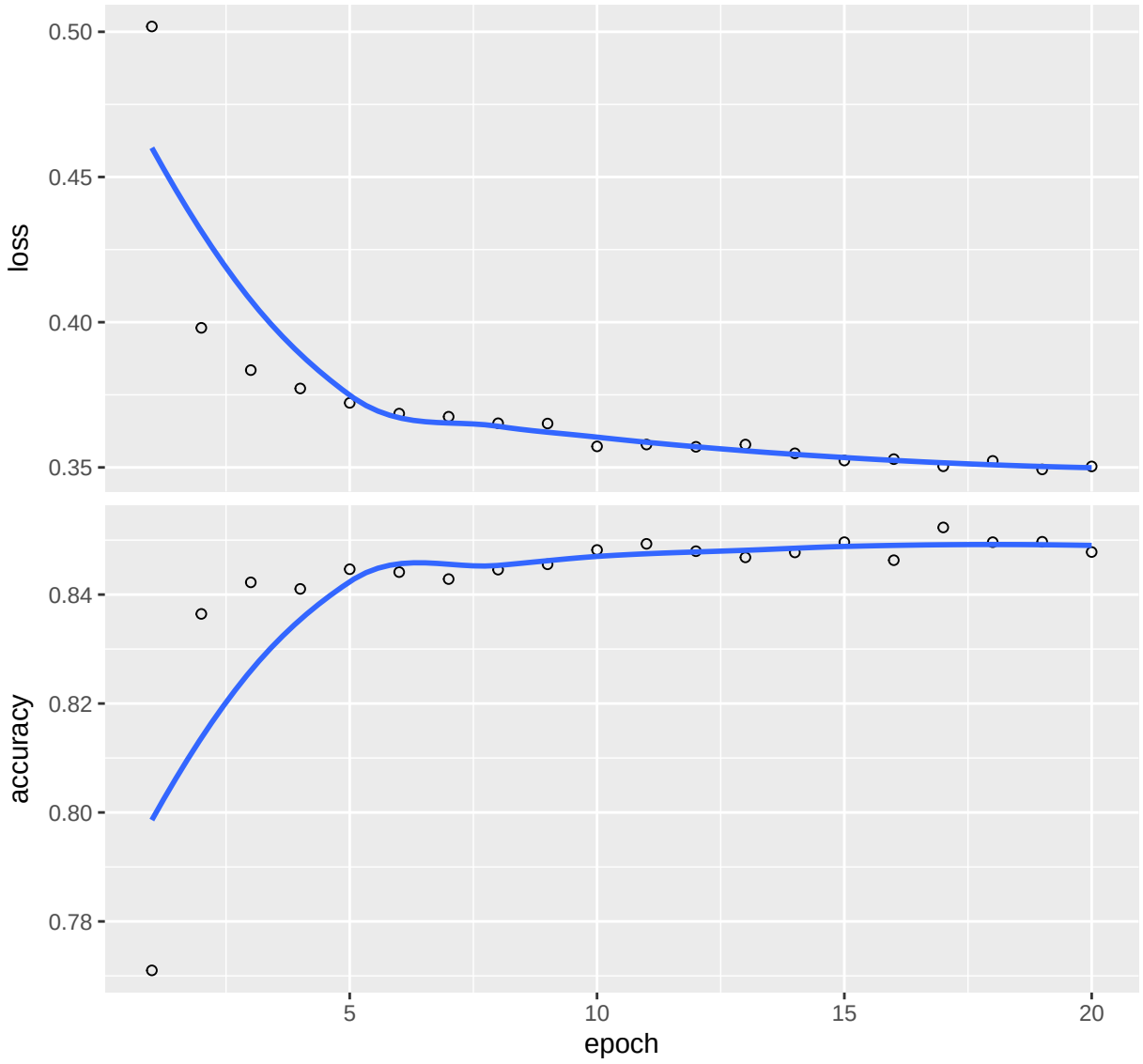
After trying some regular machine learning models, we train a multi -layer perceptron(MLP) model on our dataset to see if we can improve the prediction accuracy. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. To apply the MLP model, we transferred all factor variables into numbers. After that, we normalized everything with the same scale, which helps our model train the data more easily. I set "adam", "binary_crossentropy", and "accuracy"my optimizer, loss function, and metric. I fit the model with validation 0.2 that 20% of the train will be selected for validation, and set epochs with 100.

```
set_random_seed(43)
one_hot_train_labels <- to_categorical(train_yc)
one_hot_test_labels <- to_categorical(test_yc)
model <- keras_model_sequential() %>%
  layer_dense(units = 256, activation = "relu",input_shape = ncol(train_xc)) %>%
  layer_dropout(rate = 0.6) %>%
  layer_dense(units = 128, activation = "relu") %>%
  layer_dense(units = ncol(one_hot_train_labels), activation = "sigmoid")

model %>% compile(
  optimizer = "adam",
  loss = "binary_crossentropy",
  metrics = c("accuracy")
)

history <-model %>% fit(train_xc, one_hot_train_labels, epochs = 100,
batch_size = 512, validation_split = 0.2)
plot(history)
```



The graph above shows the loss and accuracy over 100 epochs. As we can see, the training loss decreases with every epoch, and training accuracy increases with each epoch. However, the validation loss goes down and then goes up. To avoid overfitting, we should stop training before the validation loss increases. Therefore, we will stop our model at epochs 20 (0.4345).

```
## `geom_smooth()` using formula 'y ~ x'
```


loss
0.50
0.45
0.40
0.35
accuracy
0.84
0.82
0.80
0.78
5
10
15
20
epoch

Model Evaluation

Logistic

The confusion matrix is shown below:

	Actual 0	Actual 1
Predicted 0	1384	426
Predicted 1	471	3416

0 and 1 represent patients who have survived less than 24 months or more than 24 months. The accuracy calculated based on the above confusion matrix is 84.254871%.

Xgboost

The confusion matrix is shown below:

	Actual 0	Actual 1
Predicted 0	1381	349
Predicted 1	703	3264

The accuracy calculated based on the above confusion matrix is 81.5341408%.

MLP

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4854  875
##          1  843 4822
##
##                Accuracy : 0.8492
##                  95% CI : (0.8425, 0.8557)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6984
##
##  Mcnemar's Test P-Value : 0.4545
##
##             Sensitivity : 0.8520
##             Specificity : 0.8464
##          Pos Pred Value : 0.8473
##          Neg Pred Value : 0.8512
##              Prevalence : 0.5000
##          Detection Rate : 0.4260
##    Detection Prevalence : 0.5028
##       Balanced Accuracy : 0.8492
```

```
##
##        'Positive' Class : 0
##
```

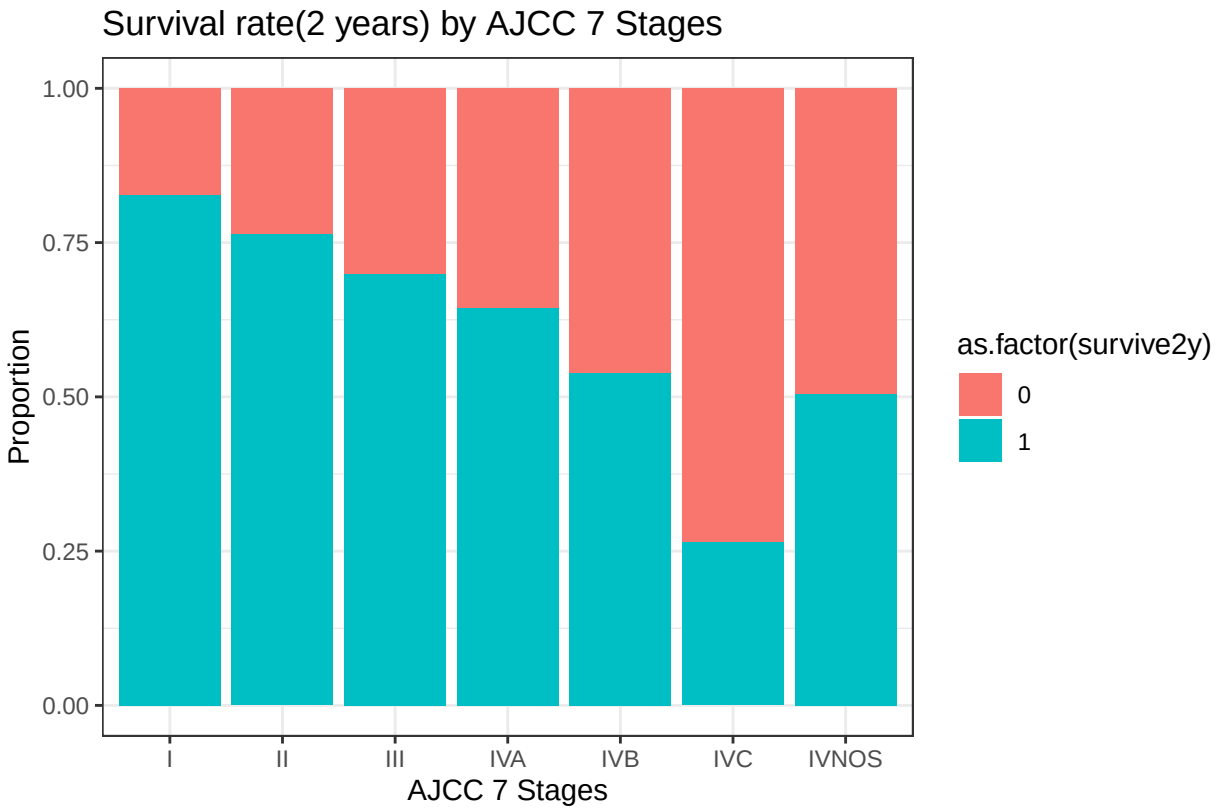
It is not easy to perfectly predict whether a patient will survive more than two years or less. Many external factors will lead to the deterioration or improvement of the cancer. In addition, the patient's internal factors, such as psychological problems, will also directly lead to the change of results. Therefore, these information not recorded in the data cannot be learned by our model. But, in general, according to our confusion matrix, the validation accuracy of 0.849 is an acceptable number, which is enough to provide doctors reference in the diagnosis of cancer.

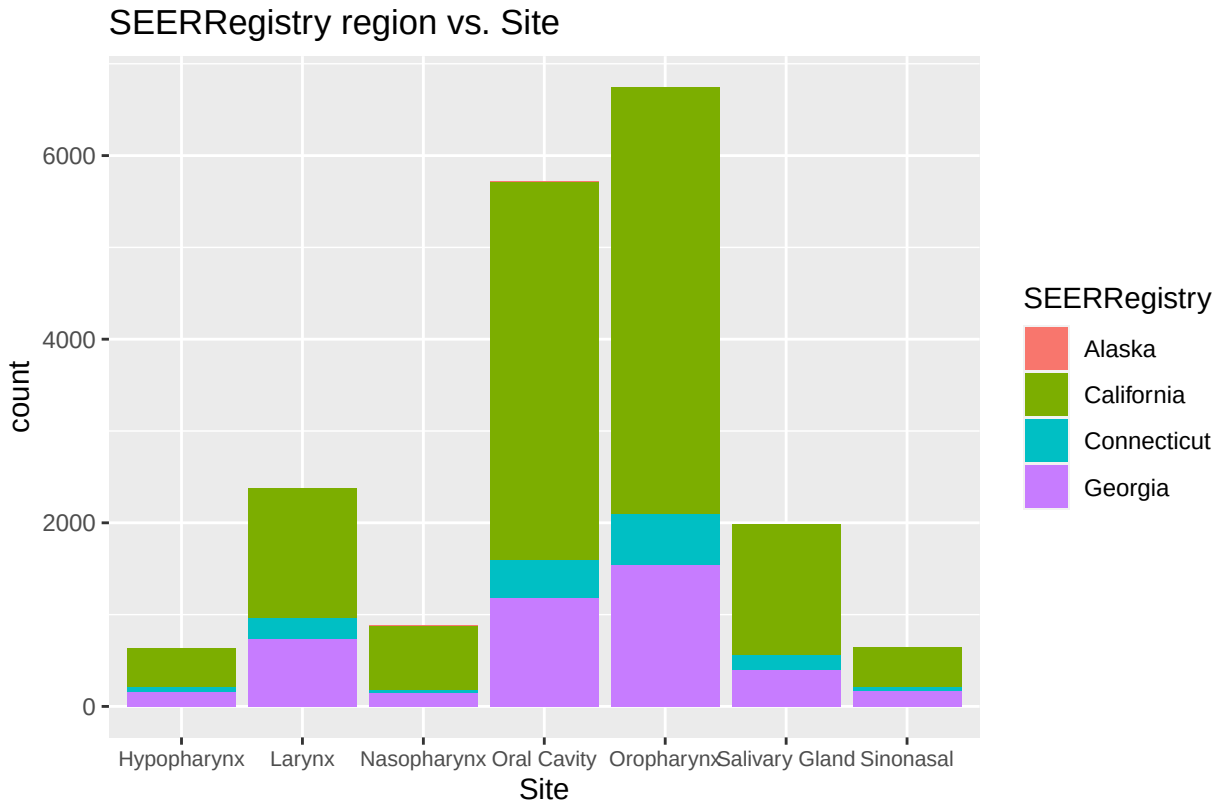
Discussion

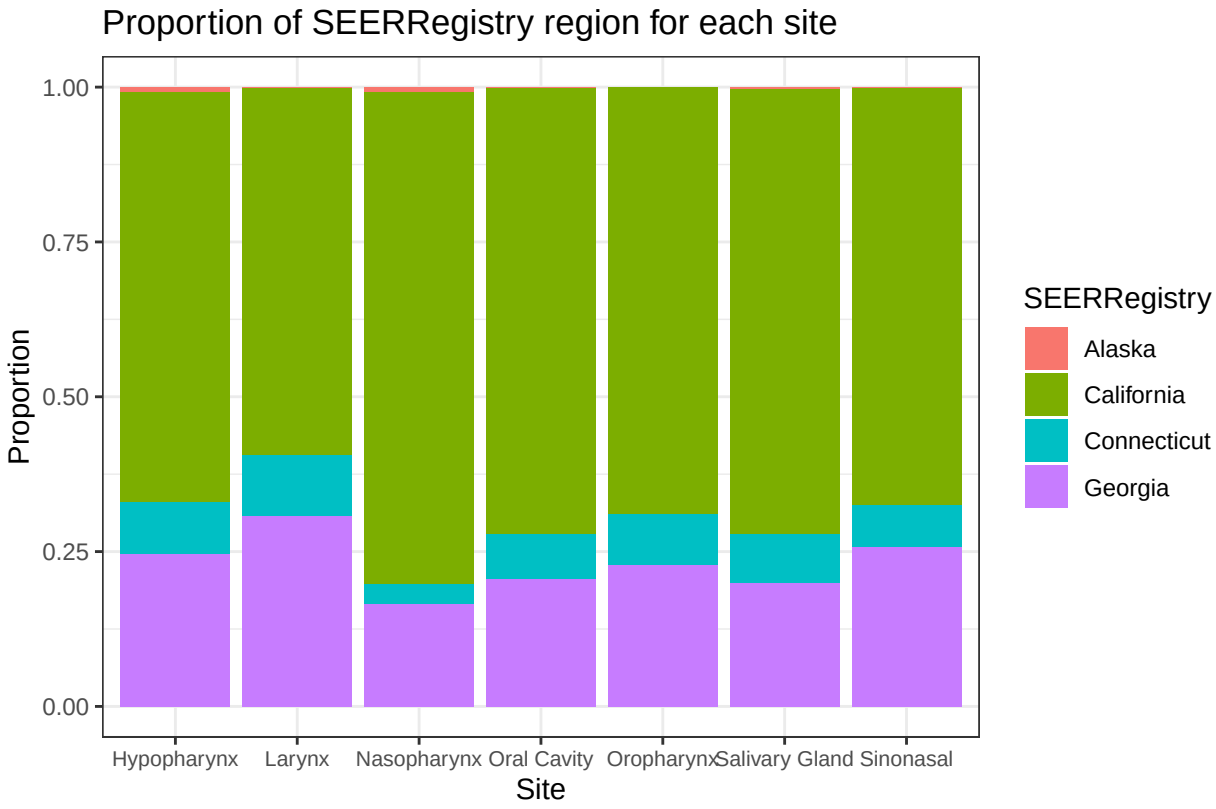
In our project, we use machine learning methods to predict whether patients with head and neck cancer can survive more than two years after diagnosis. In the end, our prediction accuracy rate reached 84.9%. We believe that this kind of prediction can help doctors estimate the patient's survival time more quickly in the early stage of diagnosis and then adjust the treatment plan, thereby improving the efficiency of the entire diagnosis and treatment. Finally, the experience of patient consultation will also be improved. However, there are still some limitations of our study. There are too many blank values in the original data. If we want to use more variables, we can only use data from 2010 to 2014. This will lead to a reduction in the amount of data we can use, which will affect the accuracy of the forecast.

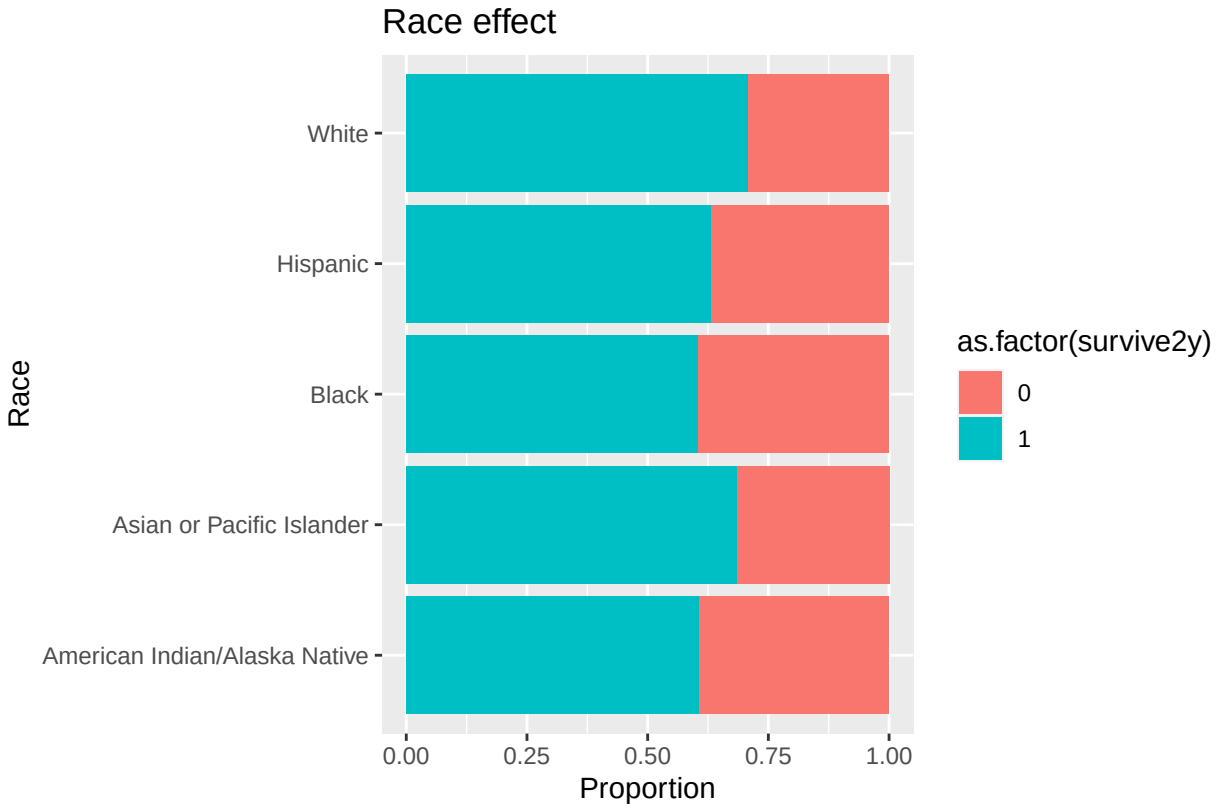
Appendix

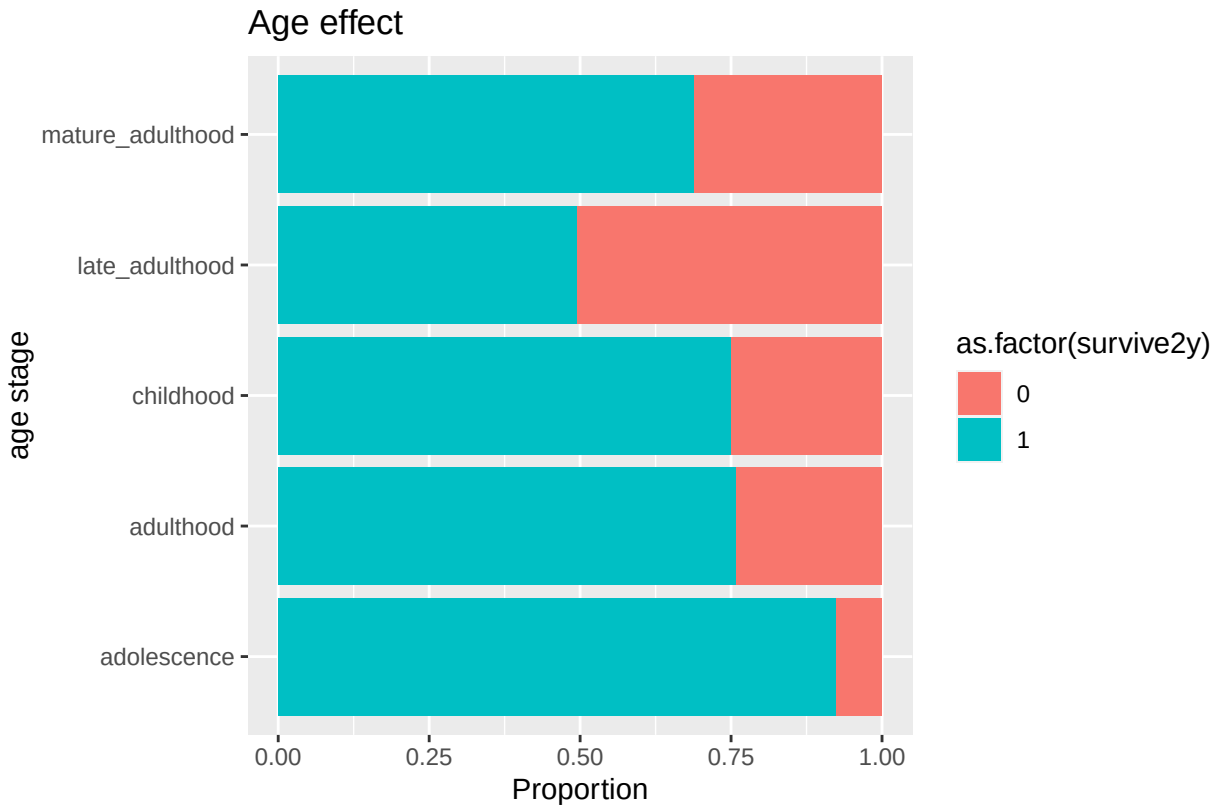
EDA

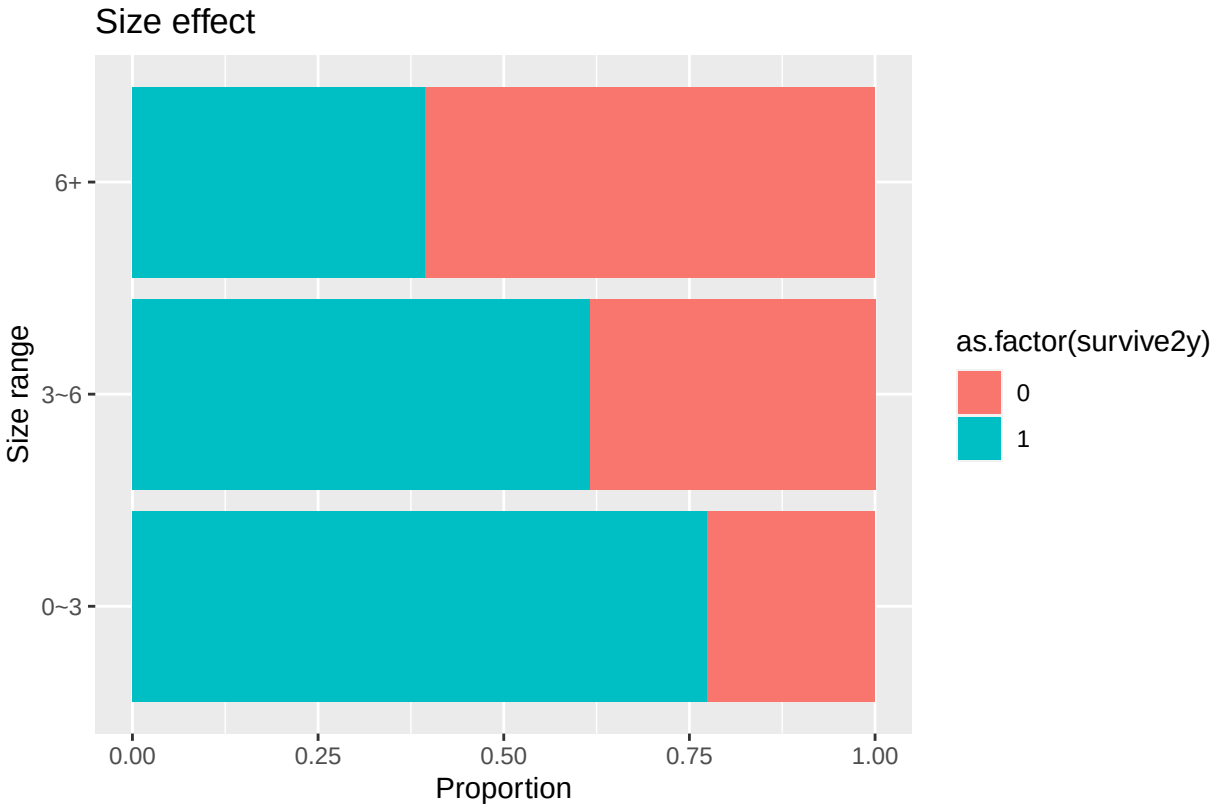

Survival rate(2 years) by AJCC 7 Stages
Proportion
1.00
0.75
0.50
0.25
0.00
I
II
III
IVA
IVB
IVC
IVNOS
AJCC 7 Stages
as.factor(survive2y)
0
1

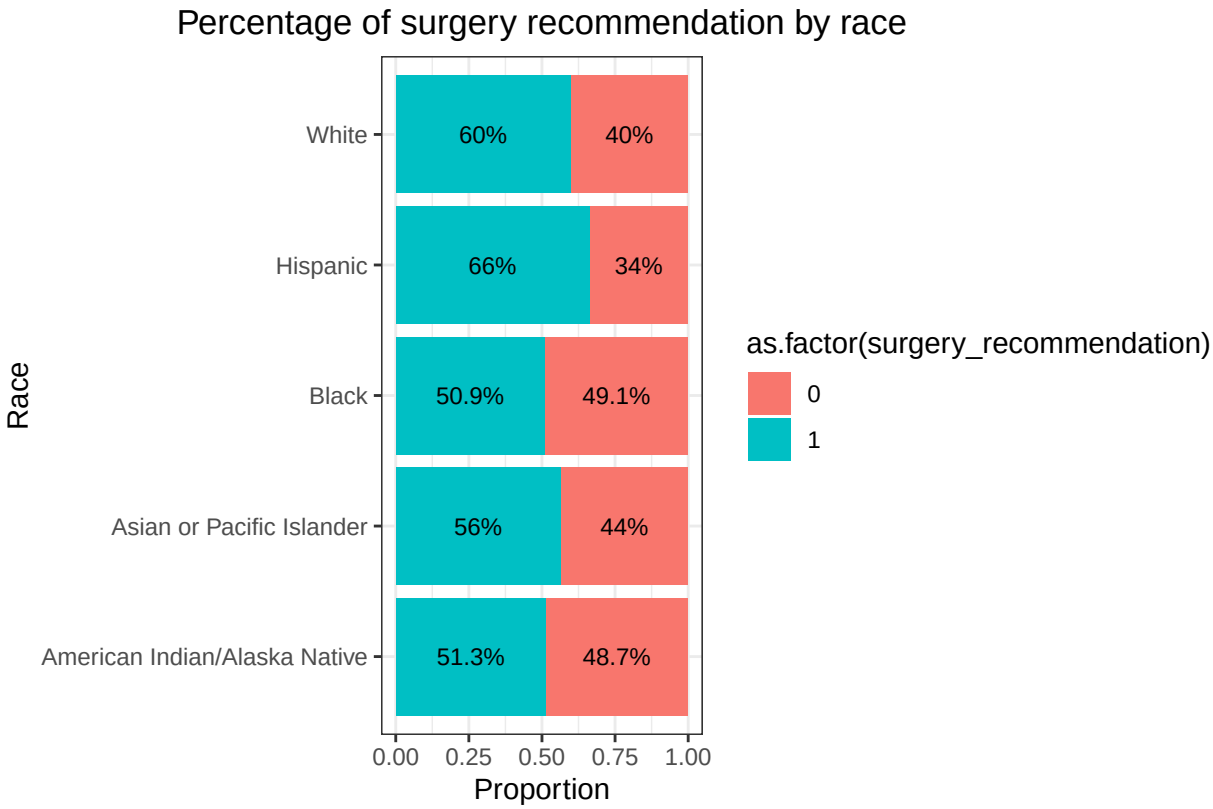

SEERRegistry region vs. Site
count
6000
4000
2000
0
Hypopharynx
Larynx
Nasopharynx
Oral Cavity
Oropharynx
Salivary Gland
Sinonasal
Site
SEERRegistry
Alaska
California
Connecticut
Georgia

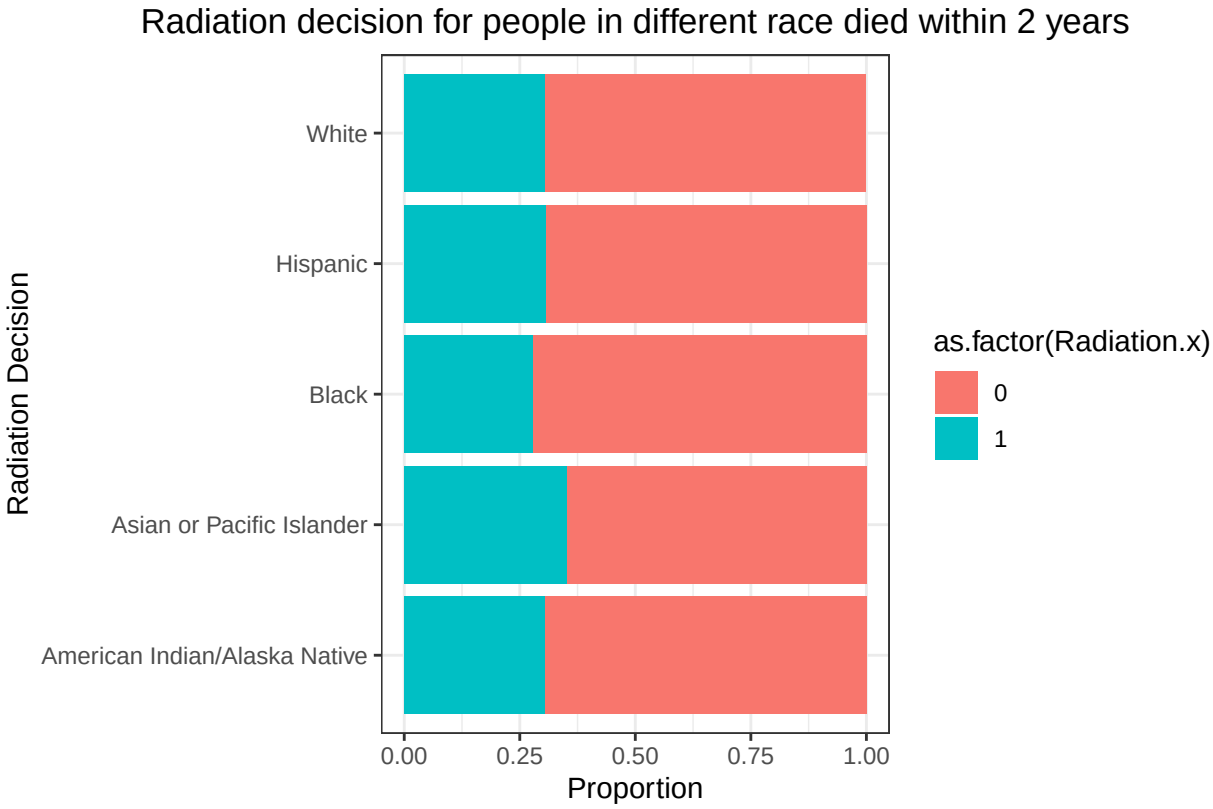

Proportion of SEERRegistry region for each site
1.00
0.75
0.50
0.25
0.00
Proportion
Hypopharynx
Larynx
Nasopharynx
Oral Cavity
Oropharynx
Salivary Gland
Sinonasal
Site
SEERRegistry
Alaska
California
Connecticut
Georgia

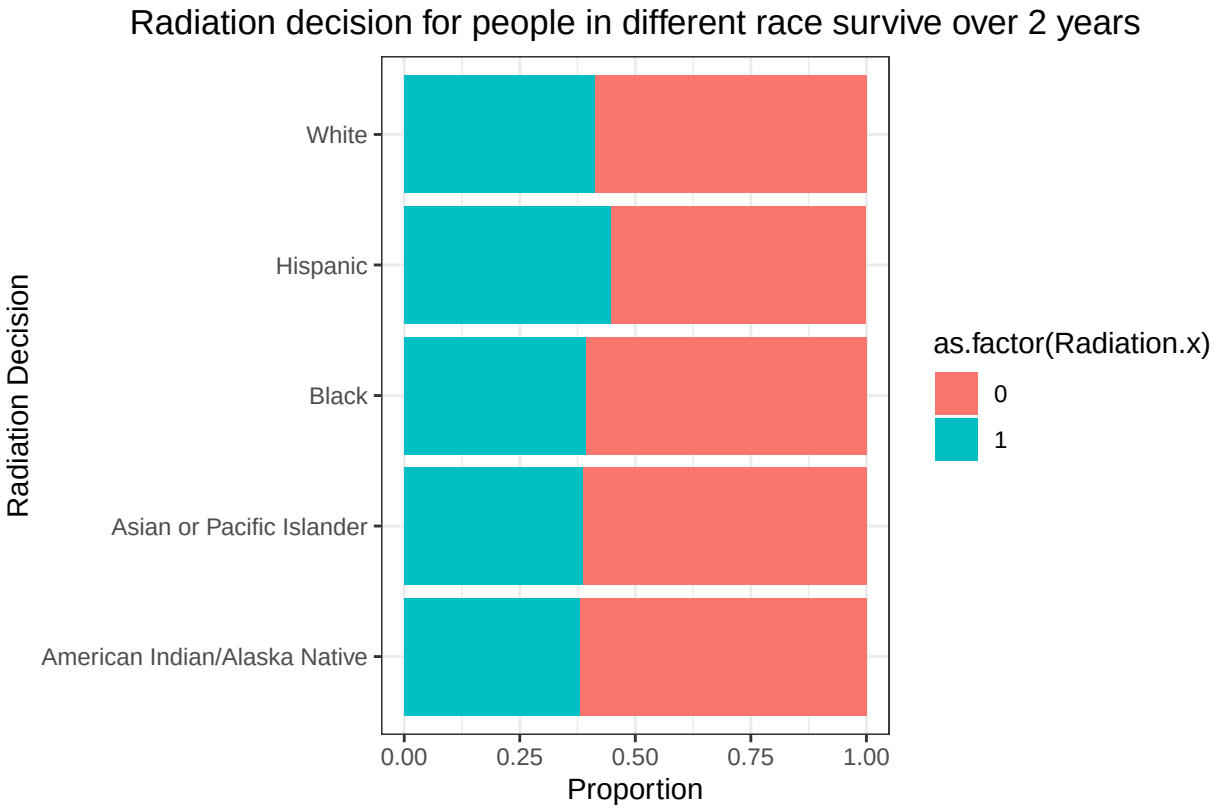

Race effect
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
Race
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(survive2y)
0
1

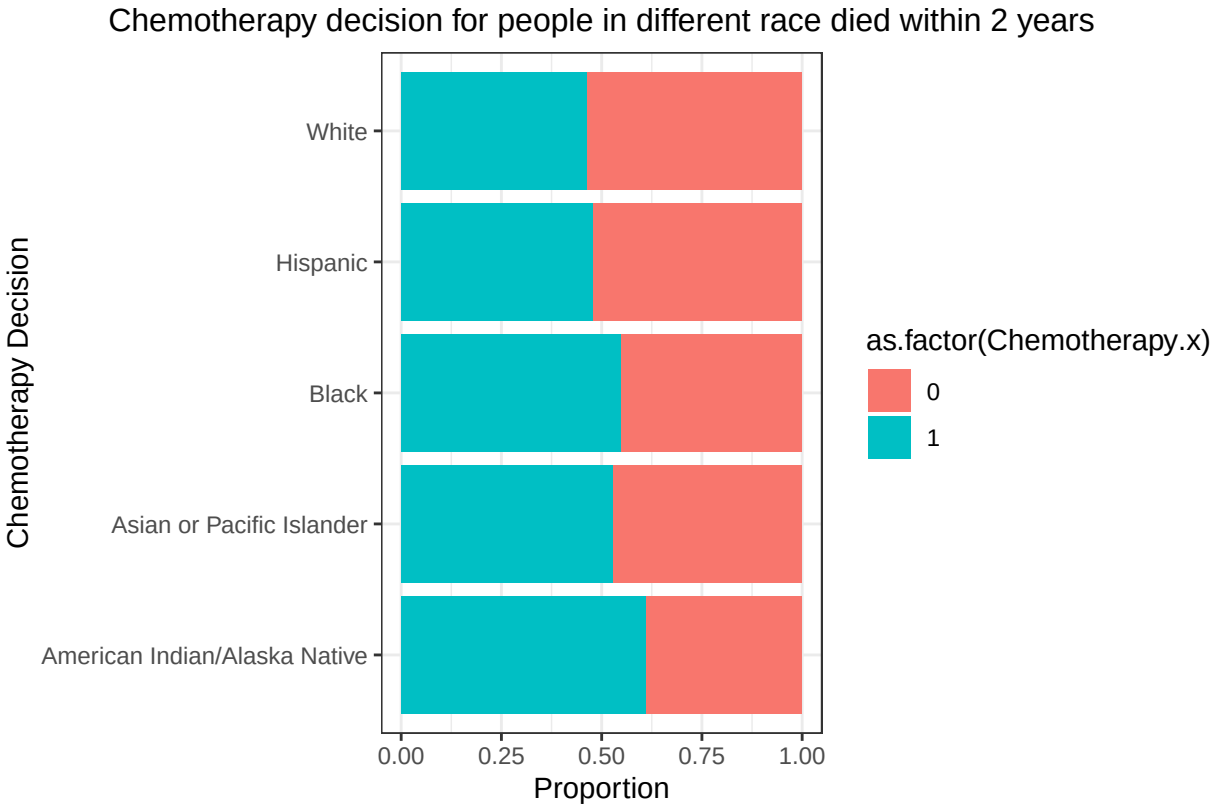

Age effect
mature_adulthood
late_adulthood
childhood
adulthood
adolescence
age stage
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(survive2y)
0
1

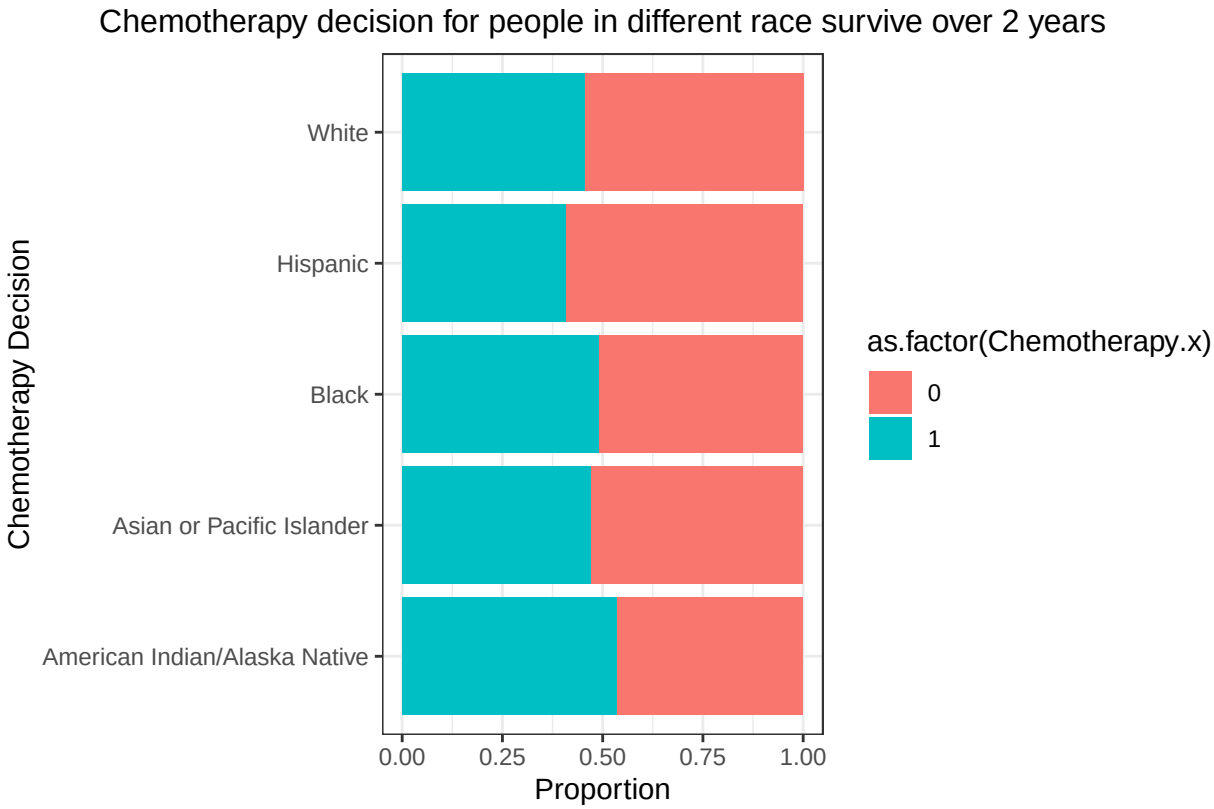

Size effect
6+
3~6
0~3
Size range
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(survive2y)
0
1

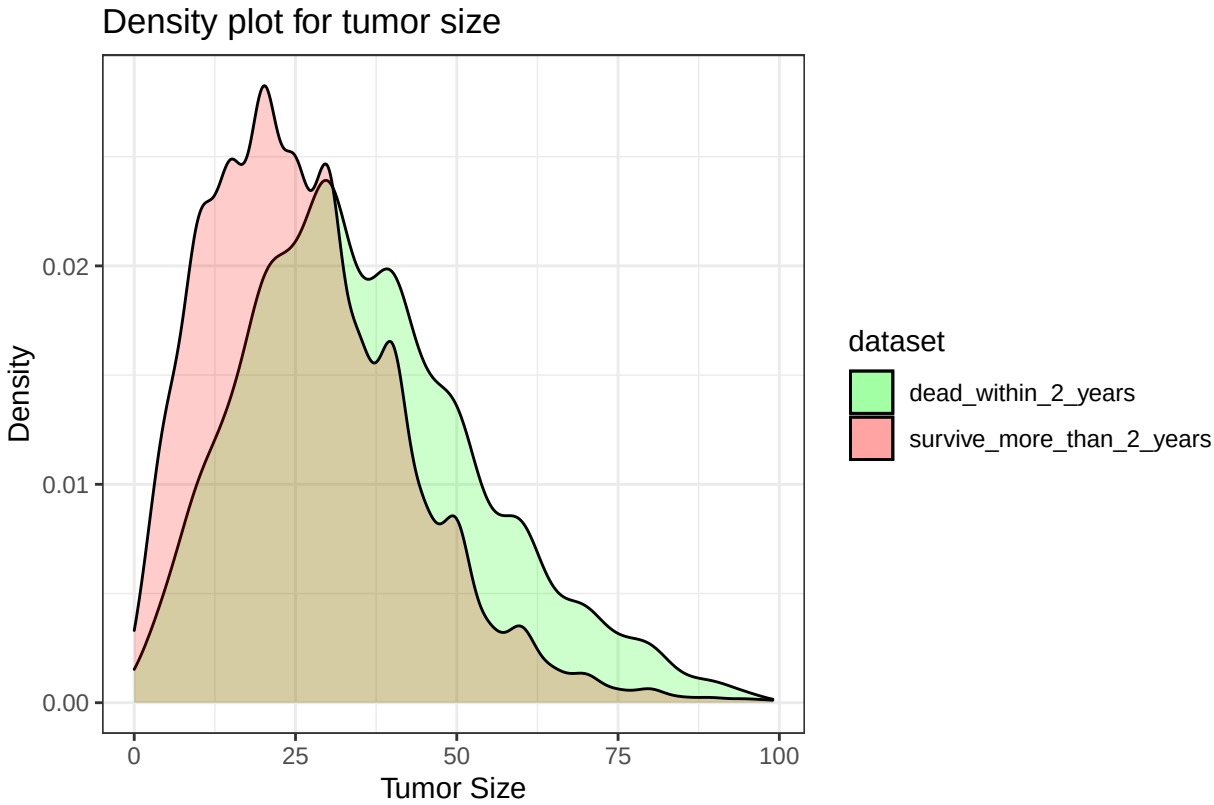

Percentage of surgery recommendation by race
Race
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
60%
40%
66%
34%
50.9%
49.1%
56%
44%
51.3%
48.7%
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(surgery_recommendation)
0
1

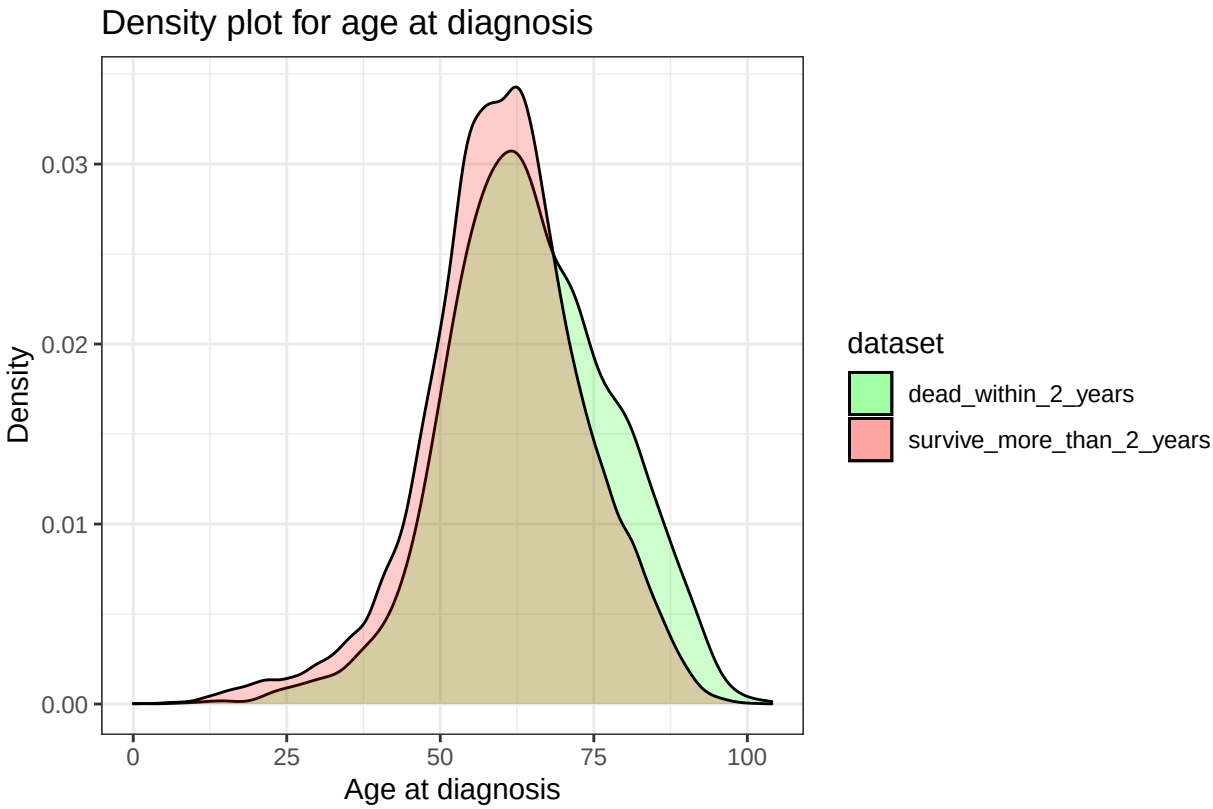

Radiation decision for people in different race died within 2 years
Radiation Decision
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(Radiation.x)
0
1

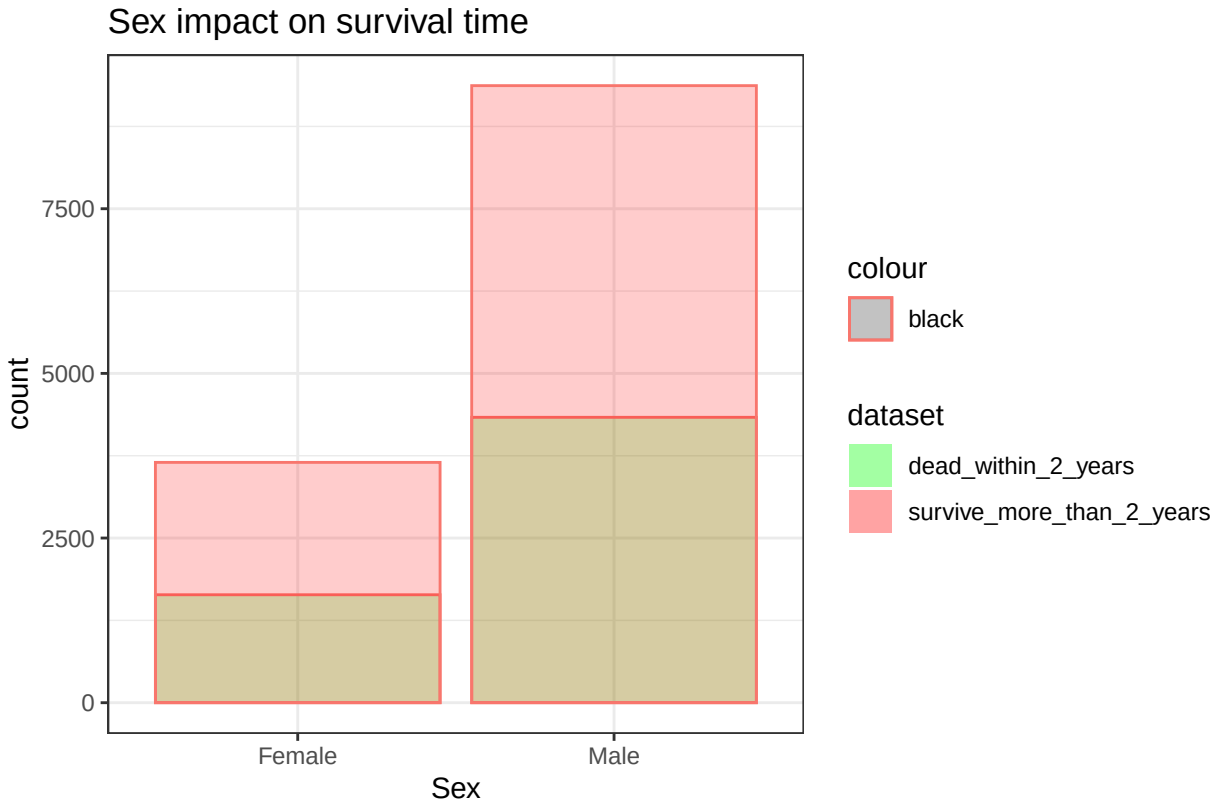

Radiation decision for people in different race survive over 2 years
Radiation Decision
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(Radiation.x)
0
1

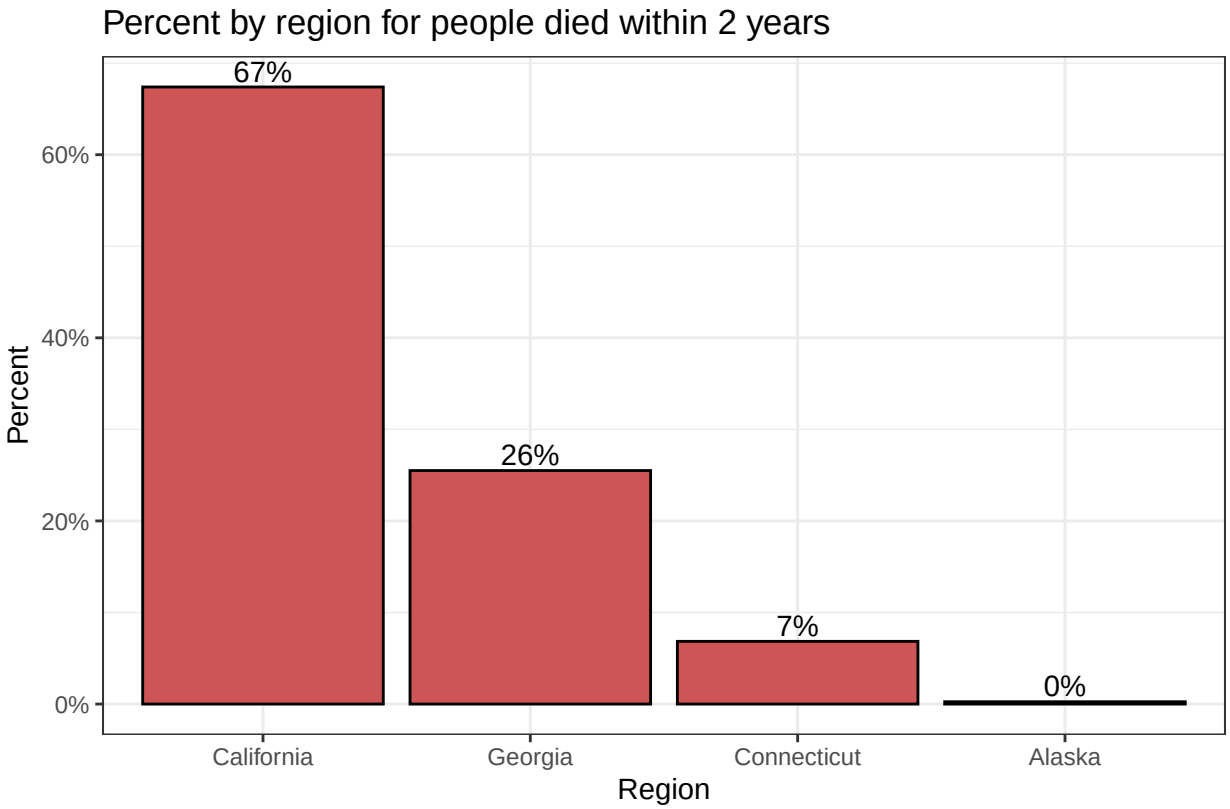

Chemotherapy decision for people in different race died within 2 years
Chemotherapy Decision
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(Chemotherapy.x)
0
1

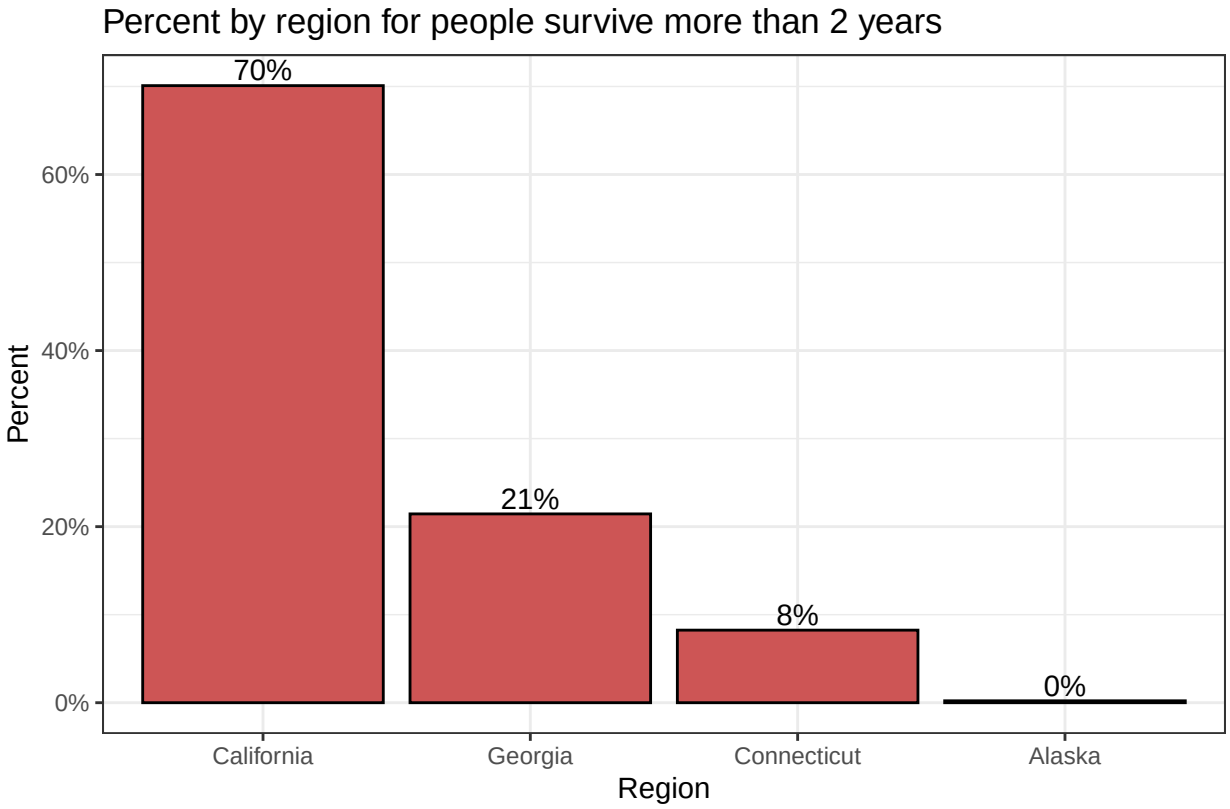

Chemotherapy decision for people in different race survive over 2 years
Chemotherapy Decision
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(Chemotherapy.x)
0
1


Density plot for tumor size
Density
0.02
0.01
0.00
0
25
50
75
100
Tumor Size
dataset
dead_within_2_years
survive_more_than_2_years

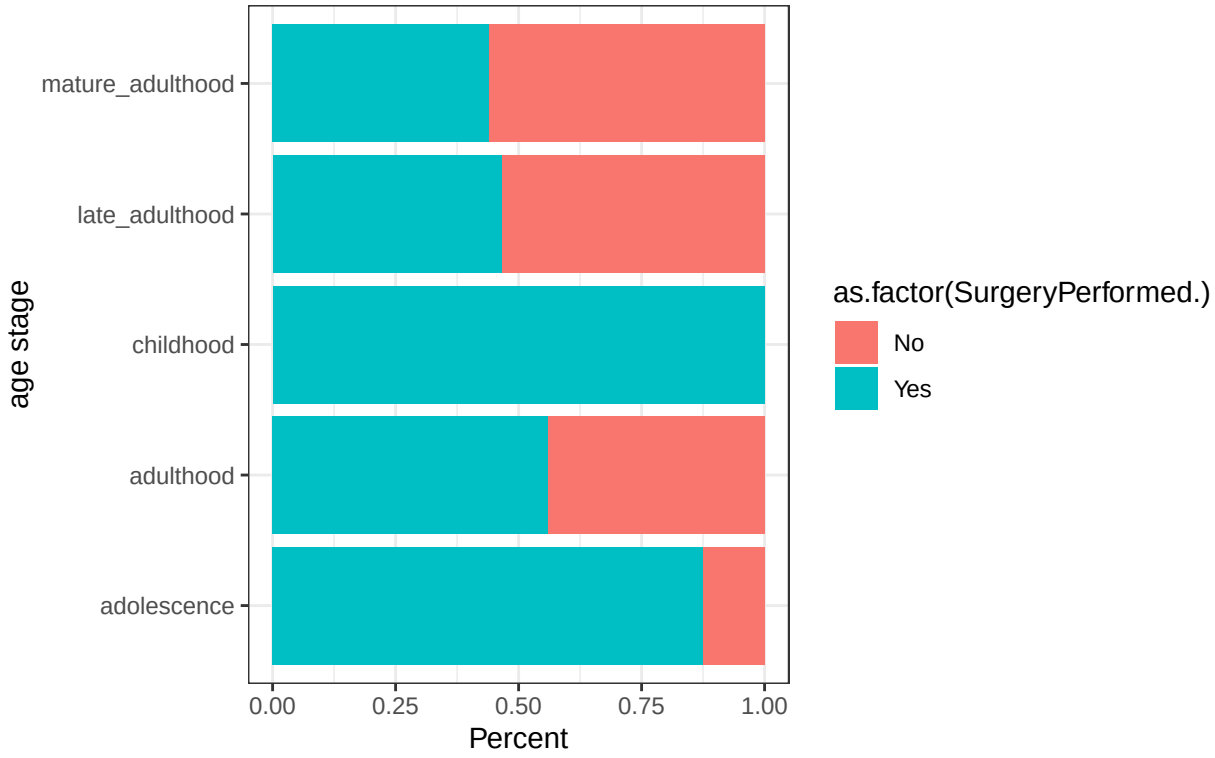

Density plot for age at diagnosis
Density
0.03
0.02
0.01
0.00
0
25
50
75
100
Age at diagnosis
dataset
dead_within_2_years
survive_more_than_2_years


Sex impact on survival time
count
7500
5000
2500
0
Female
Male
Sex
colour
black
dataset
dead_within_2_years
survive_more_than_2_years

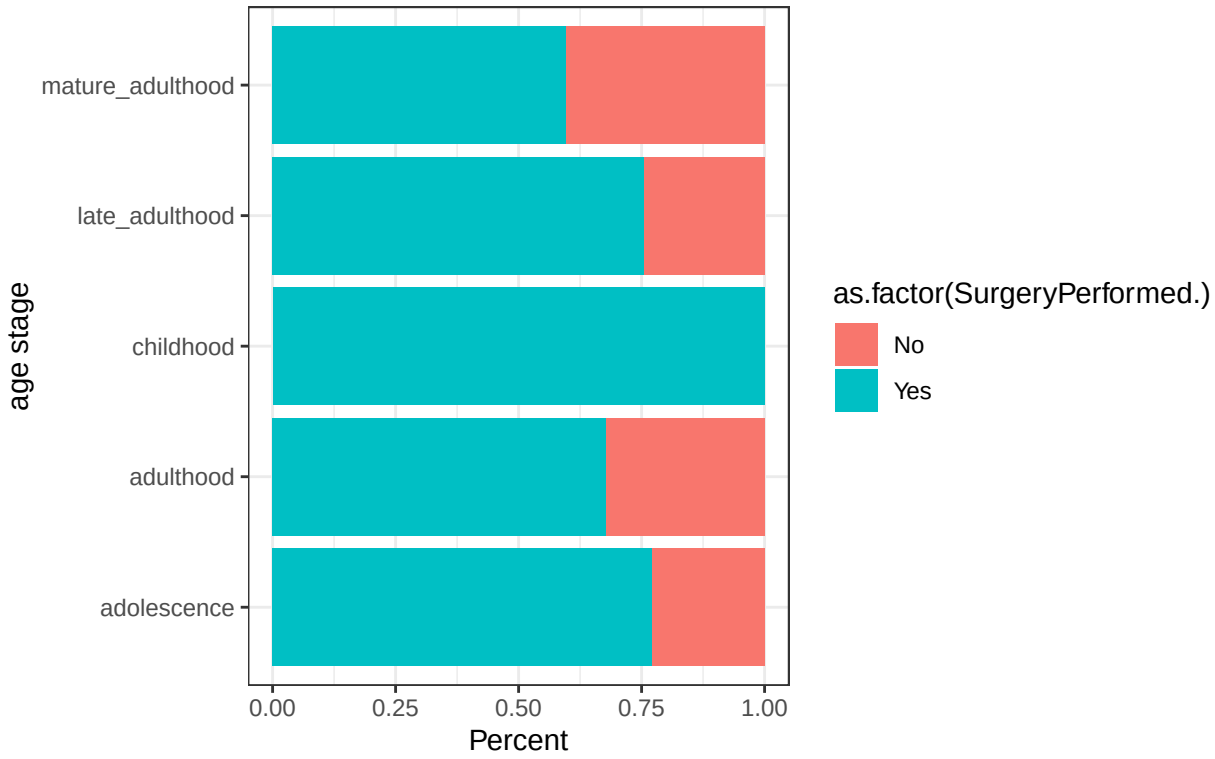

Percent by region for people died within 2 years
67%
26%
7%
0%
60%
40%
20%
0%
Percent
California
Georgia
Connecticut
Alaska
Region

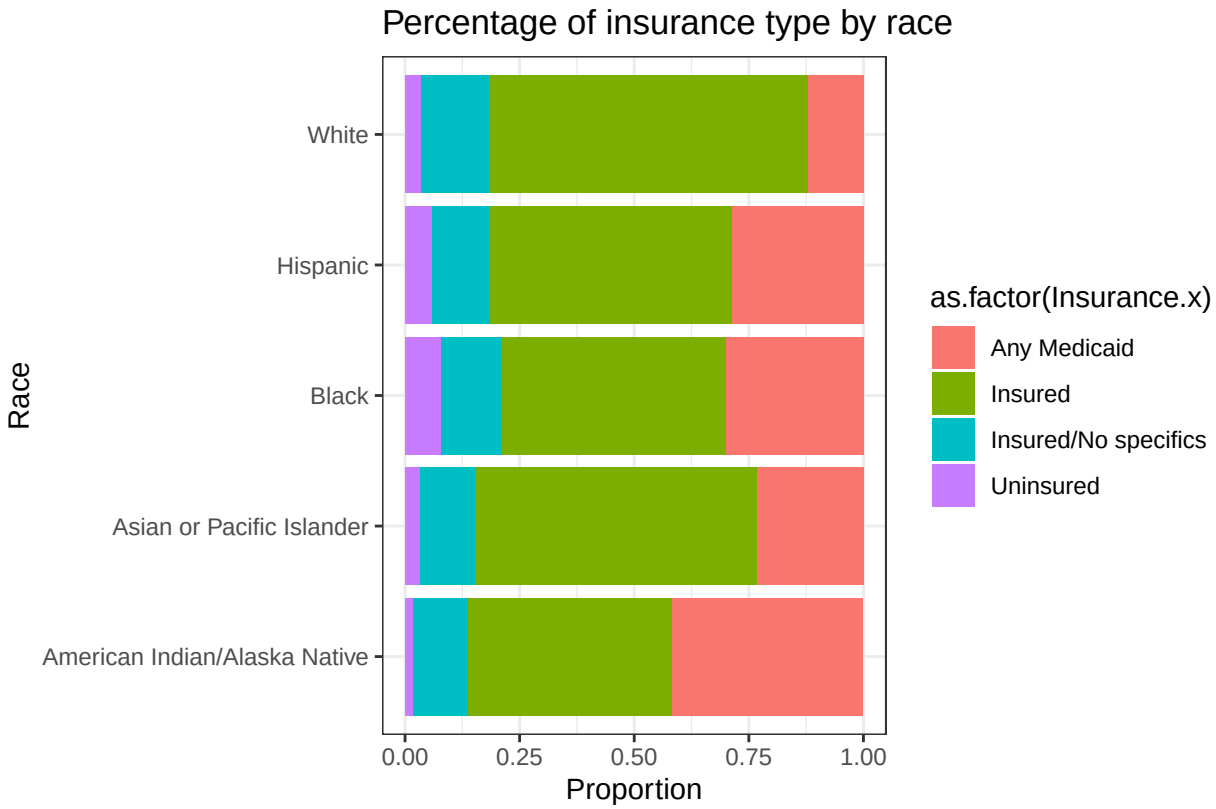

Percent by region for people survive more than 2 years
70%
21%
8%
0%
60%
40%
20%
0%
Percent
California
Georgia
Connecticut
Alaska
Region

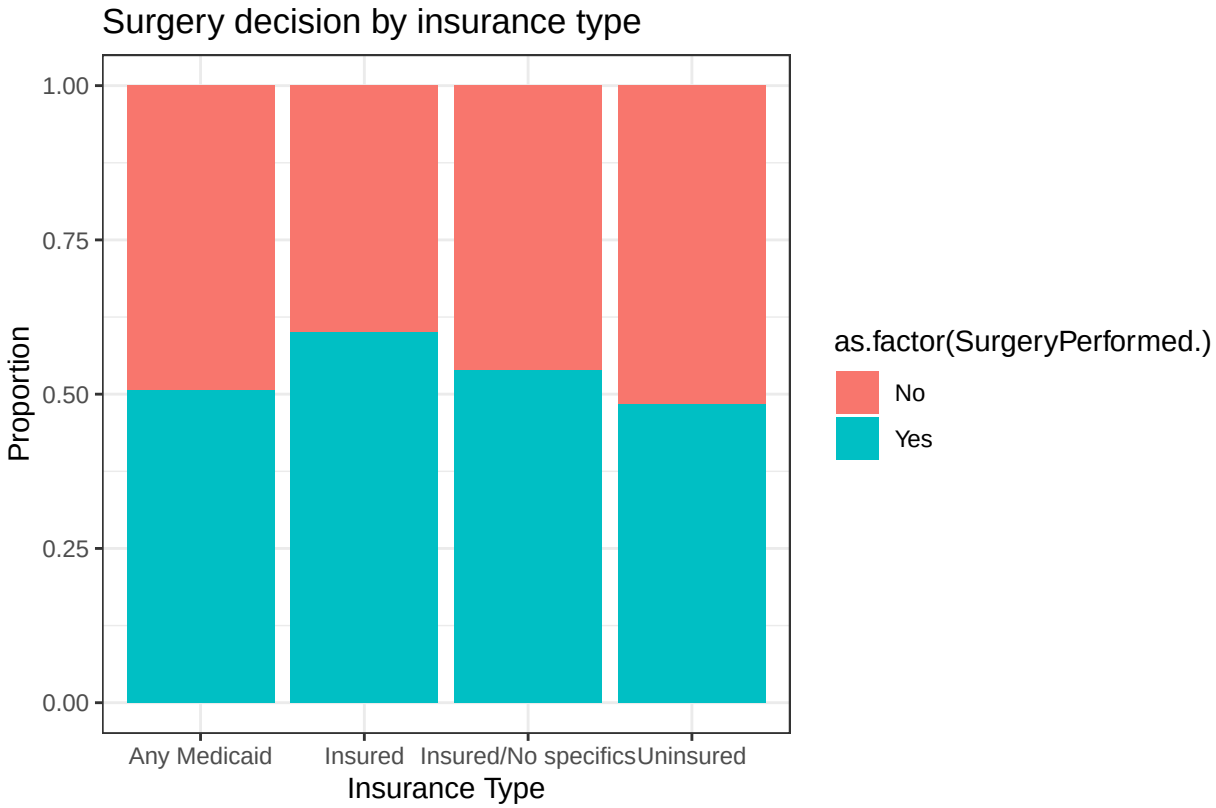
Surgery decision for people in different age stage died within 2 years

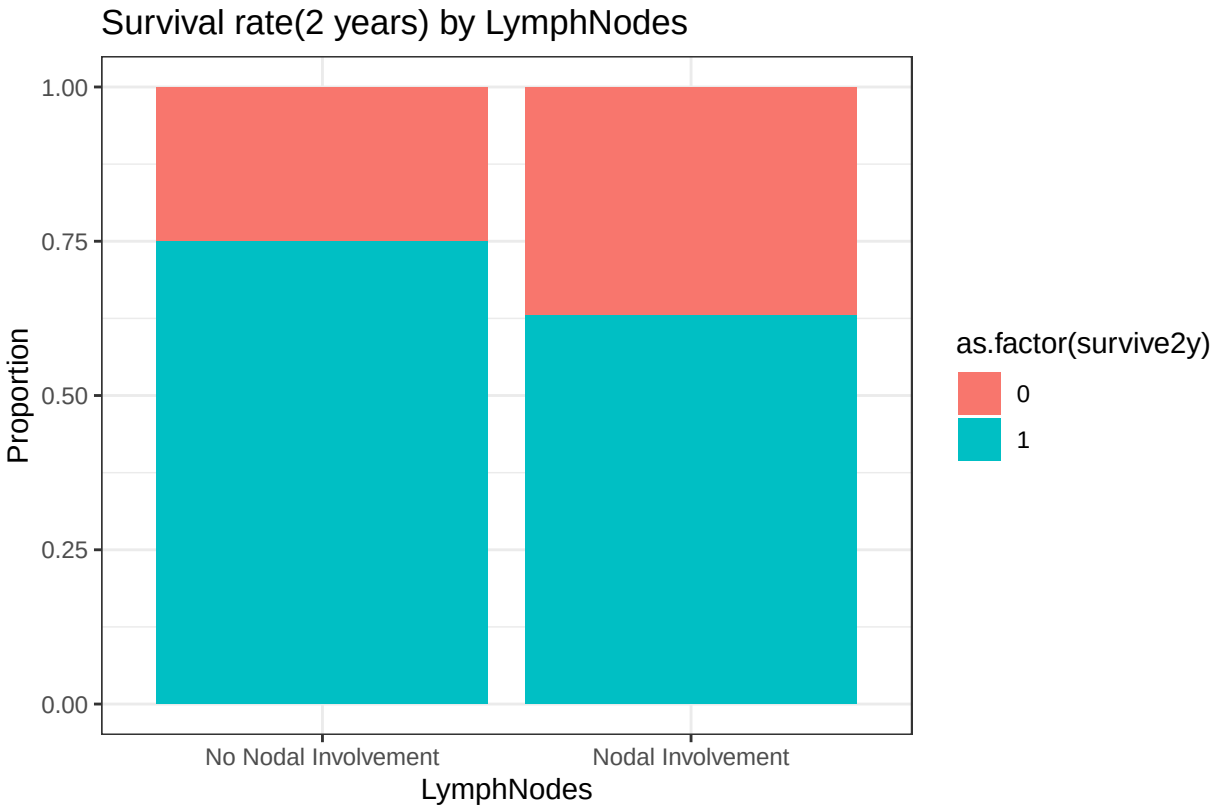


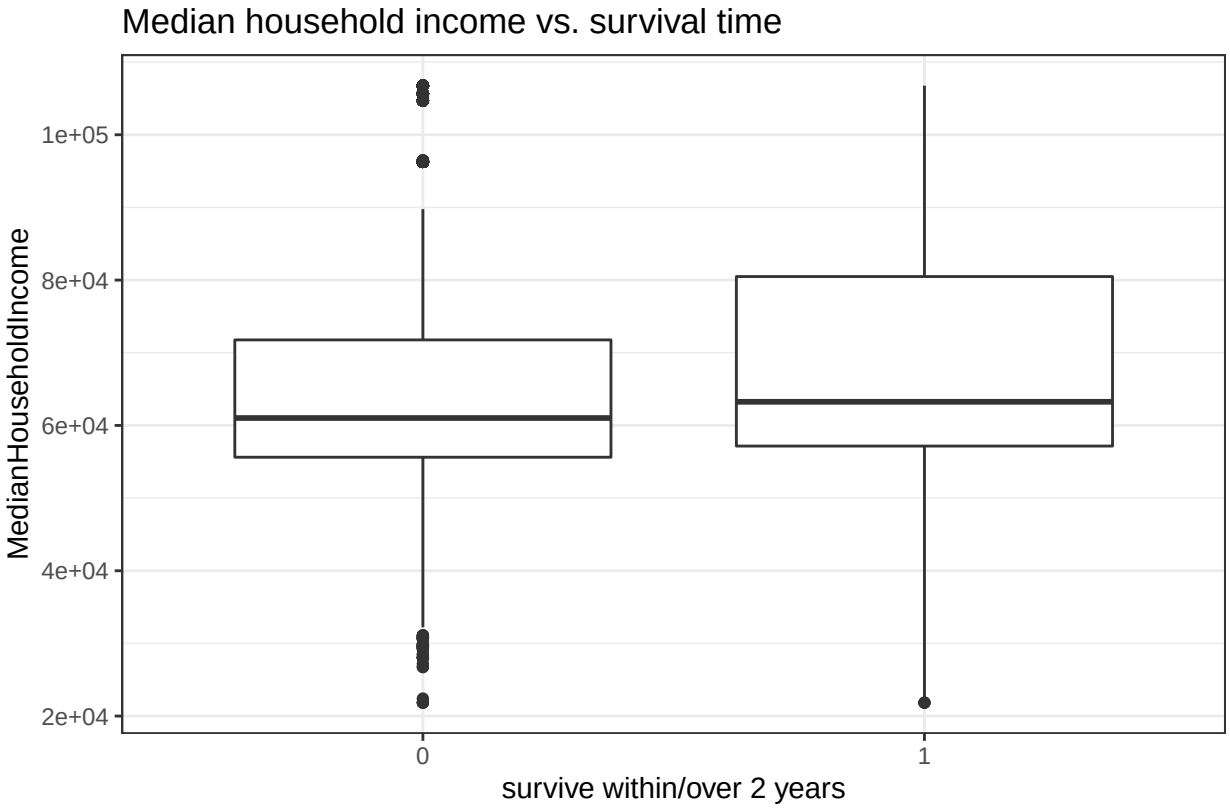
Surgery decision for people in different age stage survive over 2 years




Percentage of insurance type by race
White
Hispanic
Black
Asian or Pacific Islander
American Indian/Alaska Native
Race
0.00
0.25
0.50
0.75
1.00
Proportion
as.factor(Insurance.x)
Any Medicaid
Insured
Insured/No specifics
Uninsured


Surgery decision by insurance type
1.00
0.75
0.50
0.25
0.00
Proportion
Any Medicaid
Insured
Insured/No specifics
Uninsured
Insurance Type
as.factor(SurgeryPerformed.)
No
Yes


Survival rate(2 years) by LymphNodes
1.00
0.75
0.50
0.25
0.00
Proportion
No Nodal Involvement
Nodal Involvement
LymphNodes
as.factor(survive2y)
0
1


Median household income vs. survival time
1e+05
8e+04
6e+04
4e+04
2e+04
MedianHouseholdIncome
0
1
survive within/over 2 years