679 Final Project

Team 9(Chenghao Meng, Chun Gao, Donghao Xue, Jingwen Xu)

5/5/2021

Introduction

The treatment and management of patients with chronic and severe illnesses have been affected by several different aspects of factors, not just their medical care condition but their social background. From a practical and data acquisition point of view, the data sets provided by our clients mainly focus on cancer research.

Our main goal of the project is to explore how biases (race, gender, education level and other factors including) will affect the matching between the treatments that a patient should have and actually have. Besides, to simplify our project we would like to focus on oral cavity cancer and set particular individuals as a reference group.

Data Description

We have a dataset with 23,291 rows and 25 columns. This dataset contains information of patients who have cancers in 7 different sites, including Oral Cavity, Sinonasal, Larynx, Salivary Gland, Oropharynx, Hypopharynx and Nasopharynx. The columns are showed as below:

```
##  [1] "Study.ID"
##  [2] "Sex"
##  [3] "Year.of.Diagnosis"
##  [4] "Age.at.Diagnosis"
##  [5] "Race"
##  [6] "Insurance"
##  [7] "SEER.Registry"
##  [8] "X...9th.Grade.Education"
##  [9] "X...High.School.Education"
## [10] "X...Bachelors.Education"
## [11] "X..Persons.Below.Poverty"
## [12] "X..Unemployed.ACS.2013.2017"
## [13] "Median.Household.Income"
## [14] "X..Language.isolation.ACS.2013.2017..households."
## [15] "Site"
## [16] "Subsite"
## [17] "AJCC.7.Stage"
## [18] "Size"
## [19] "Lymph.Nodes"
## [20] "Mets"
## [21] "Cause.of.Death"
## [22] "Surgery.Performed."
## [23] "Surgery.Decision"
## [24] "Radiation"
## [25] "Chemotherapy"
```

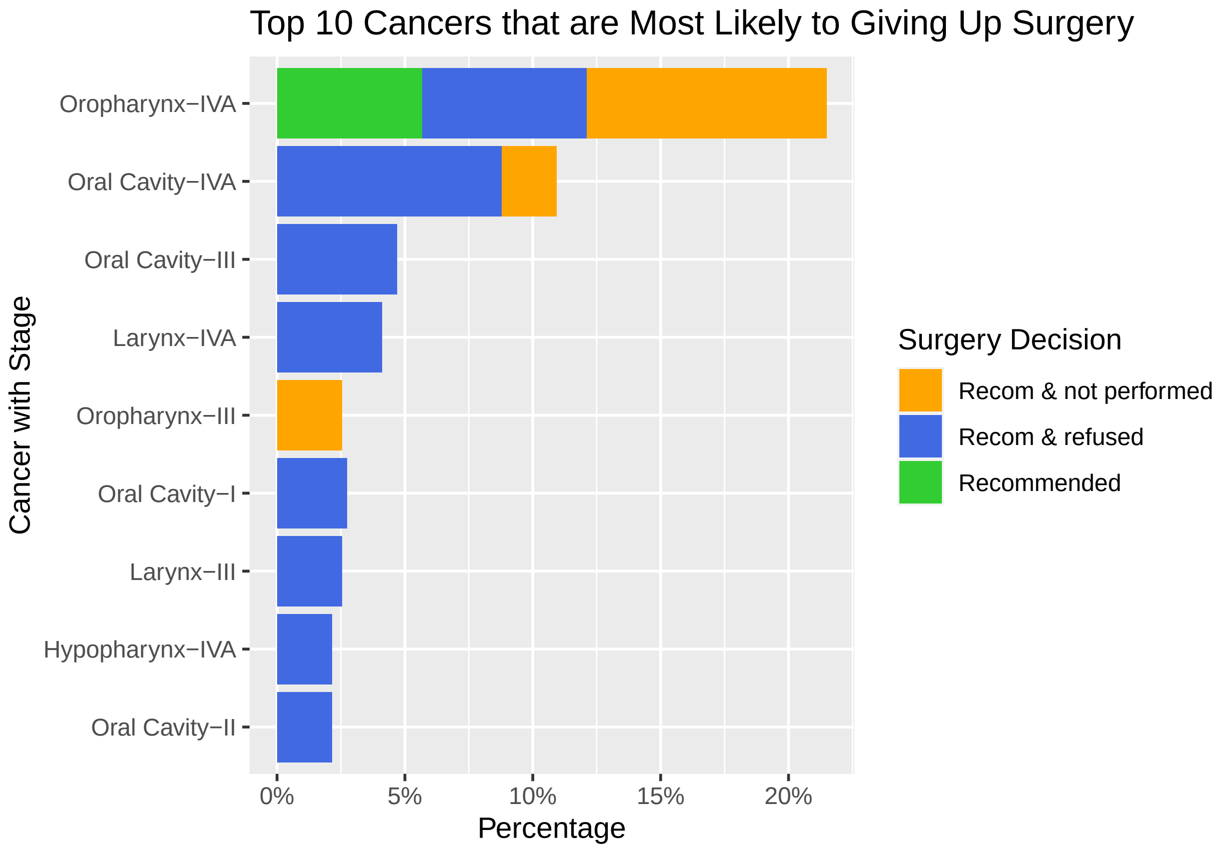
Exploratory Data Analysis

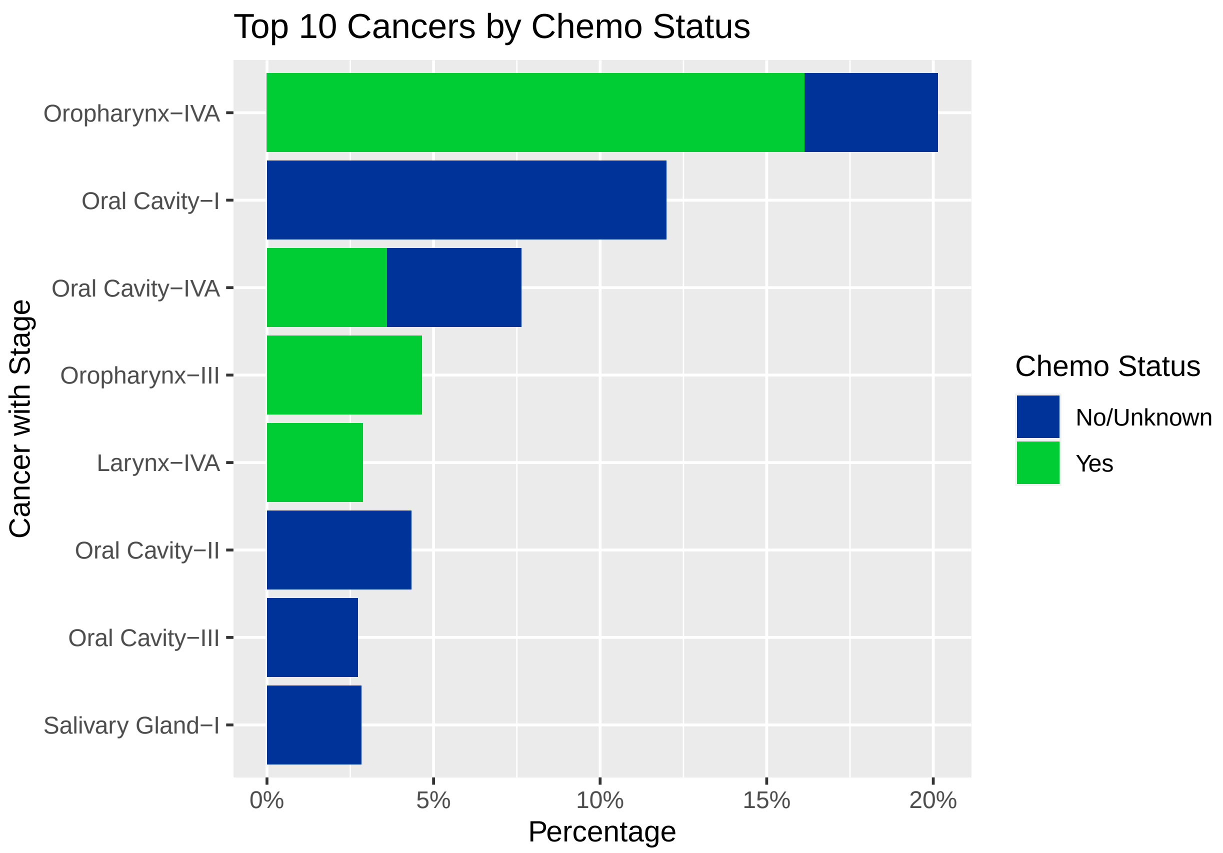
Cancer Type Selection

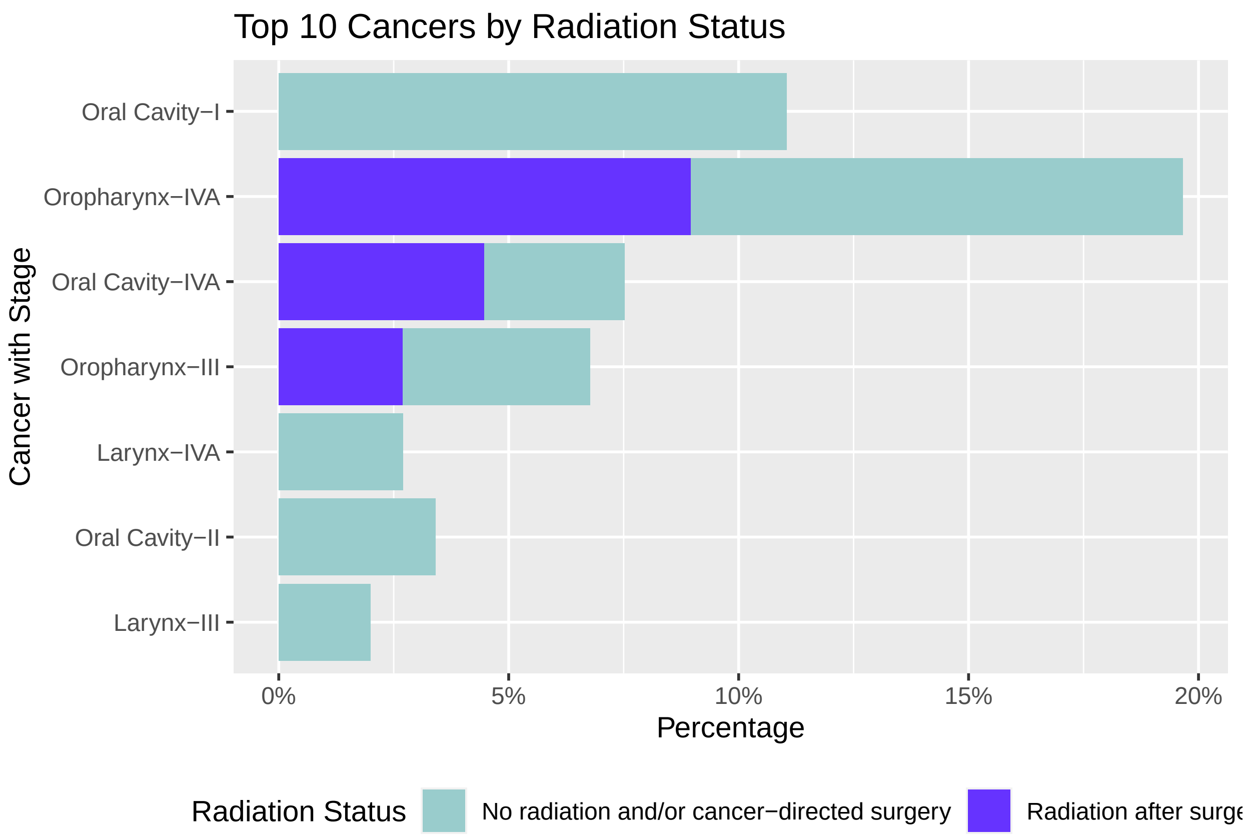
There are seven kinds of cancer with different stages in this dataset. To simplify our research, we decide to choose only one kind of cancer to do the analysis.

So, we firstly explore the top 10 cancers (including the stages) that the patients are most likely to give up the surgery. Giving up the surgery can suggest that the patients are affected by some factors such as biases in the society.

The plot below shows that, even given recommendation, some patients with oral cavity cancer in nearly all the stages, even the early stage, have given up the surgery. Combining the information given by the status of chemo and radiation therapy, we would like to choose oral cavity cancer as our research direction.




Top 10 Cancers by Chemo Status
Cancer with Stage
Oropharynx−IVA
Oral Cavity−I
Oral Cavity−IVA
Oropharynx−III
Larynx−IVA
Oral Cavity−II
Oral Cavity−III
Salivary Gland−I
0%
5%
10%
15%
20%
Percentage
Chemo Status
No/Unknown
Yes



Features by Surgery Performed Status

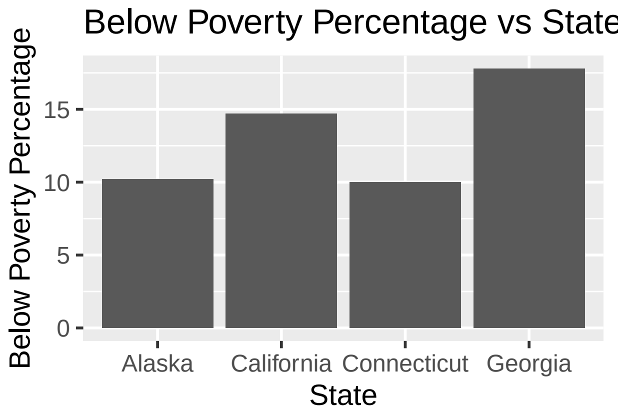
To begin with, we would like to take a look at the condition of surgery implementation.

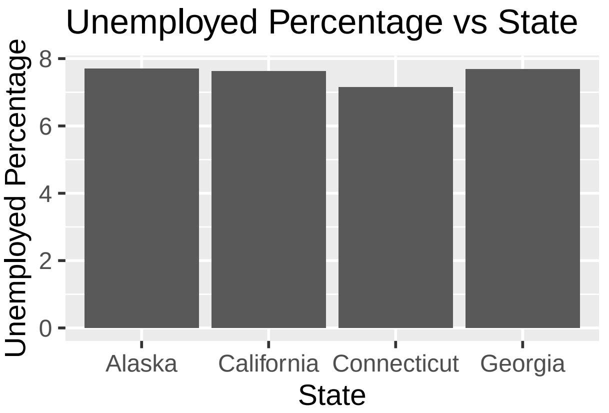
State

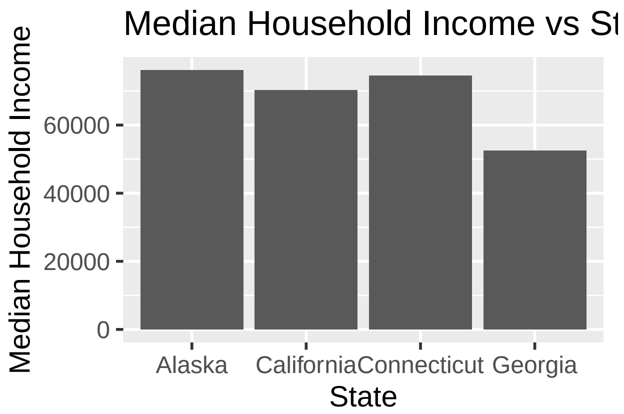
Environment Attributes by State

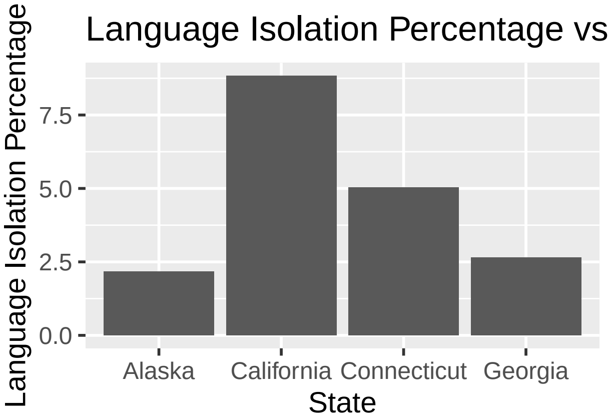
Household Income, Poverty, Unemployment, Language Isolation and Education are all recorded on county level, which can also be regarded as important attributes of the respondents' `SEER.Registry`, so we would like to explore their distribution by states.

For Poverty, the state of Georgia has the highest poverty rate, over 15% in average. For Unemployment, the unemployment rates are quite similar across those 4 states. For Household Income, the state of Georgia has the lowest household income in average. For Language Isolation, the state of Alaska has the lowest language isolation rate, while the state of California has the highest.

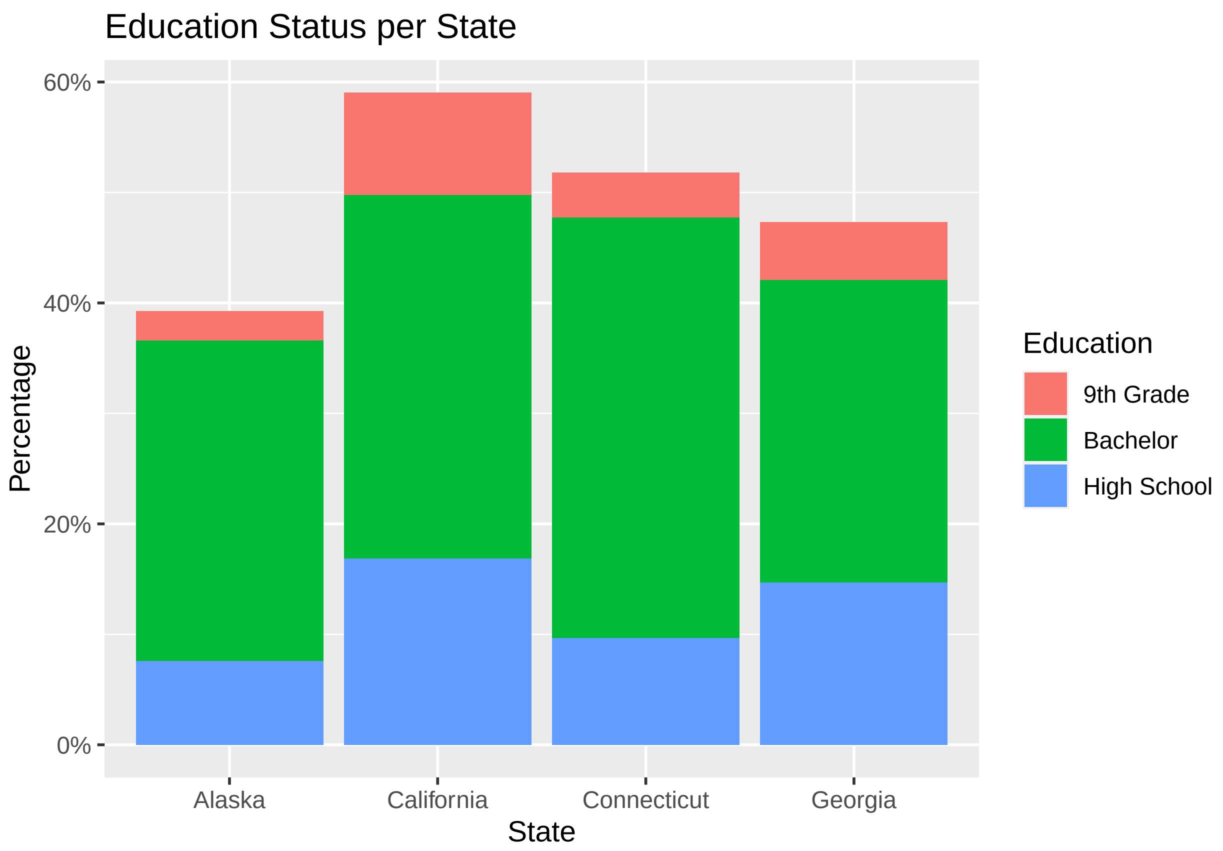




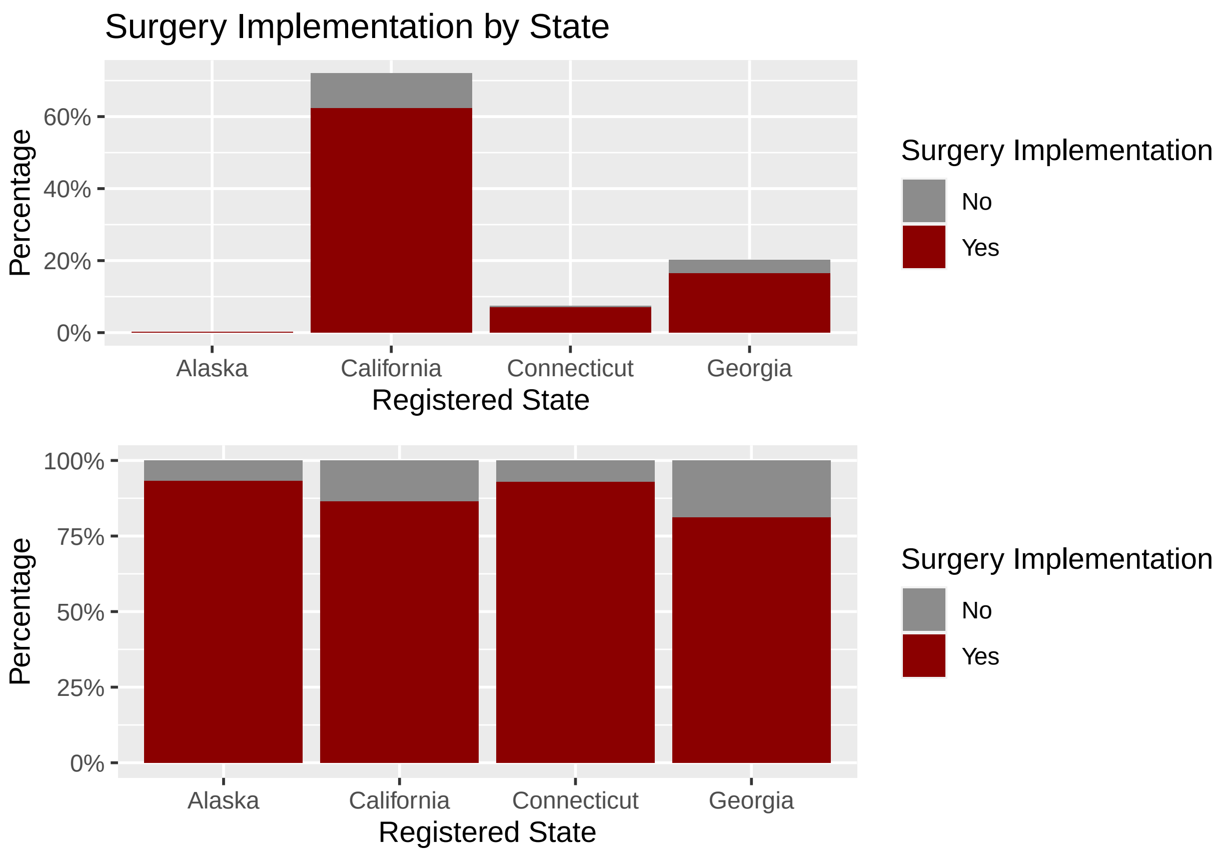




For Education, the state of California has the highest proportion of respondents are educated below 9th Grade and the highest proportion of respondents are educated below High School. Meanwhile, the state of Connecticut has the highest proportion of respondents are educated below Bachelor.

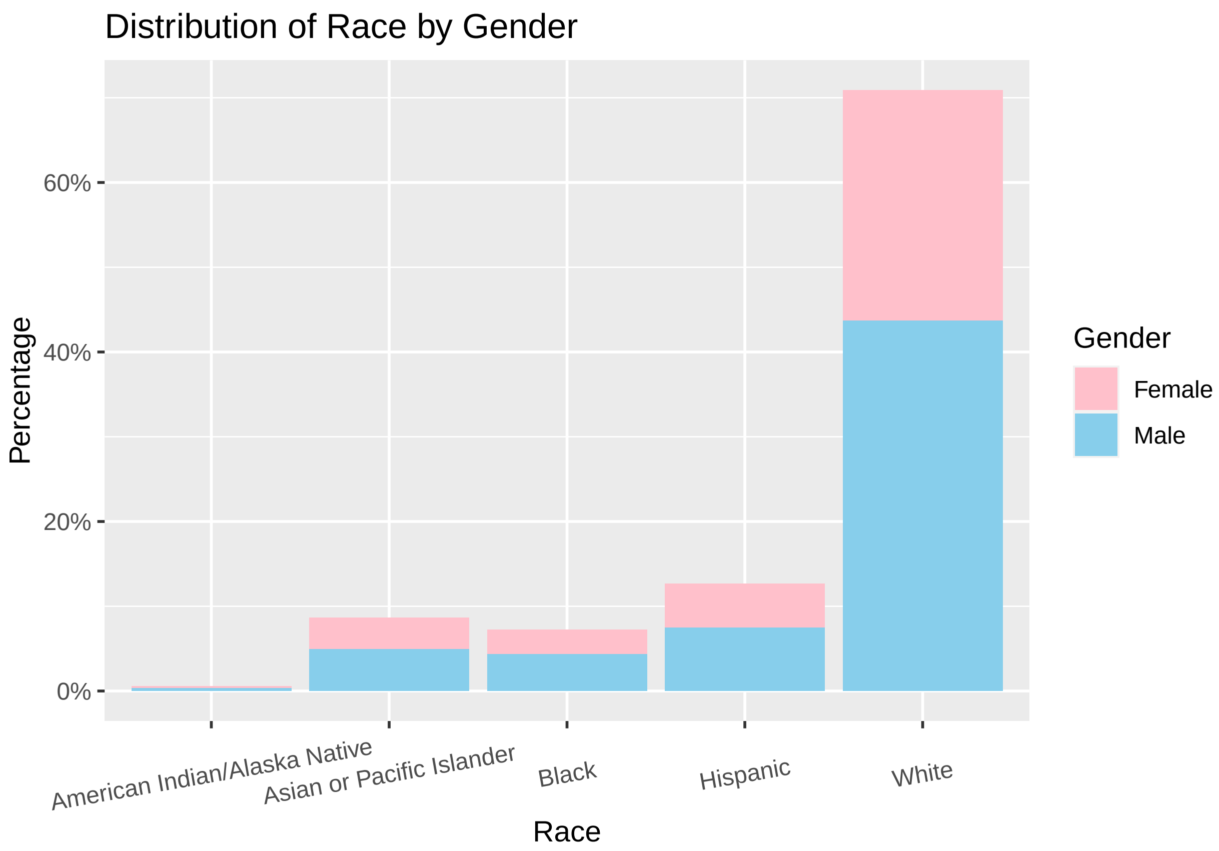


By examining surgery implementation condition by state, we found that California has the largest population of patients researched in the data. While Alaska and Connecticut have the largest proportion of patients who have performed the surgery.

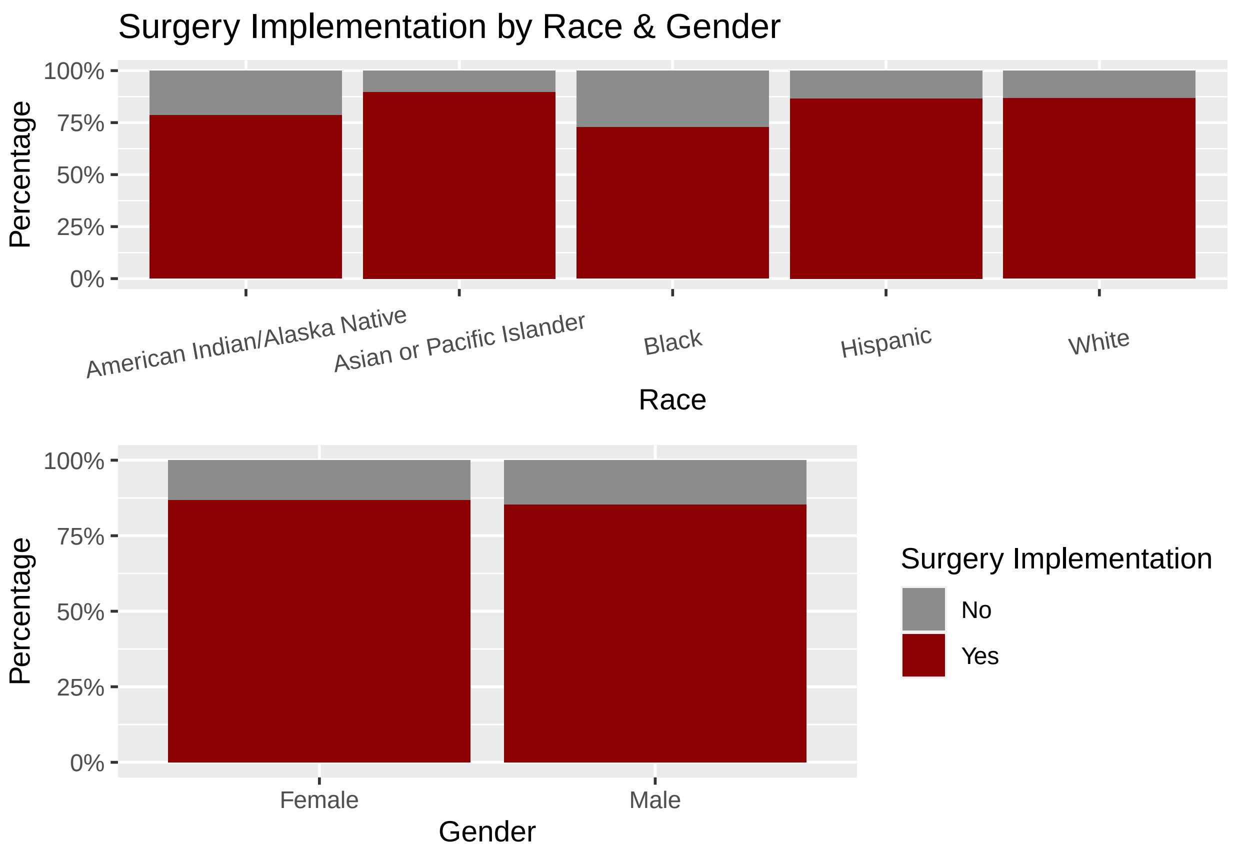


Race & Gender

Before looking at the surgery condition by race and gender, we would like to take a look at the distribution of race by gender. The plot above shows that, more than 60% of observations are white people, and more than 40% of observations are white male, which indicates that there would be an imbalance problem in the data.

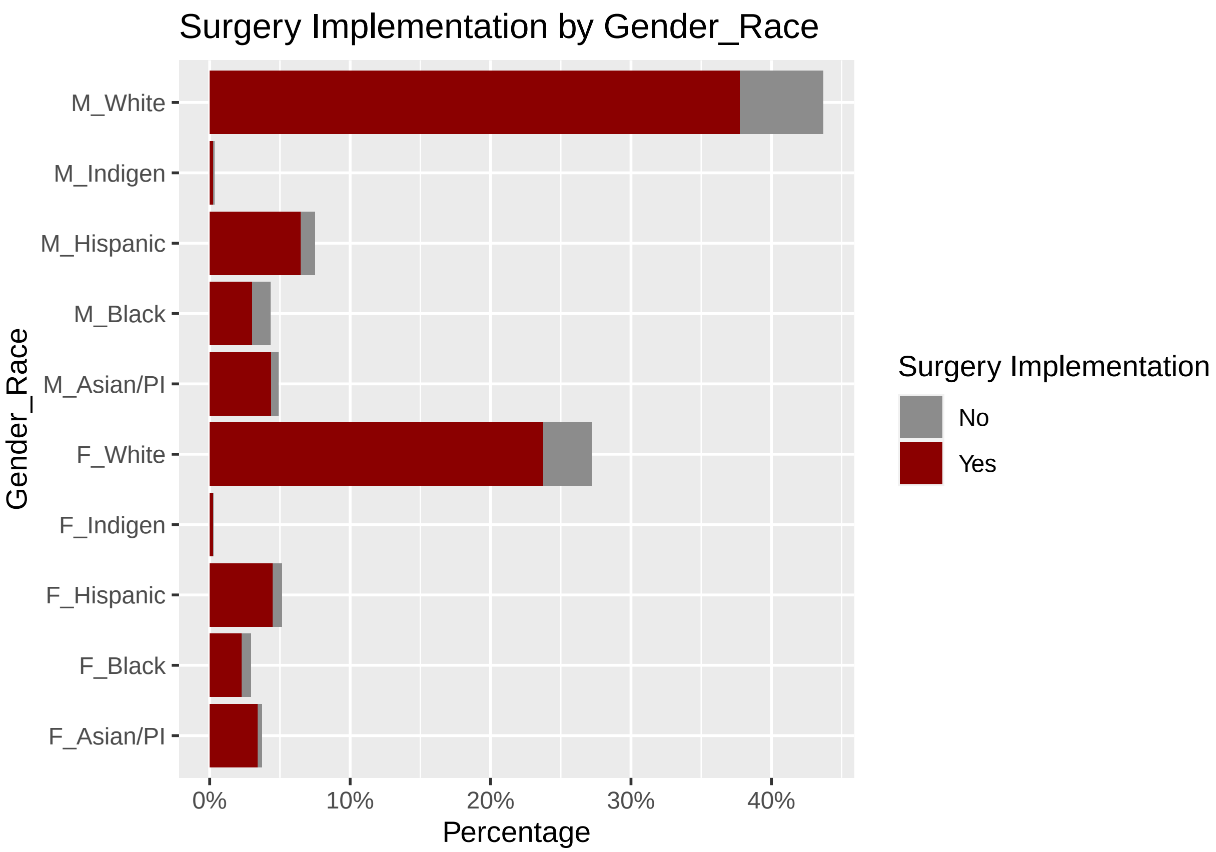


For surgery condition, the plot below shows that Asian of Pacific Islander has the largest proportion of performing the surgery while the Black has the smallest one. Female and Male has almost the same proportion to accept the surgery, which seems that there is no gender biases.



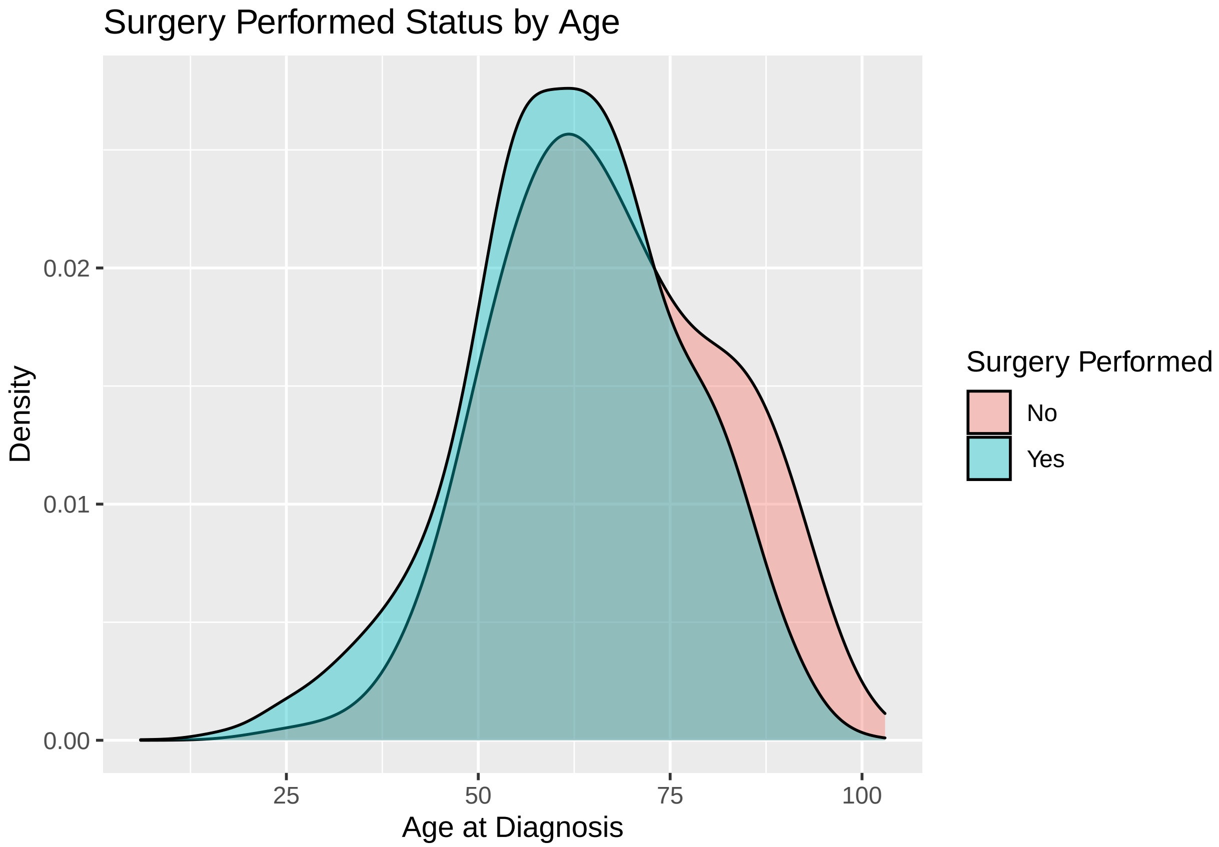
After looking at distribution by gender and race separately, we would like to put these 2 factors together to explore the distribution of performed surgery.

Then, we find that the white male has the largest proportion of performing a surgery.

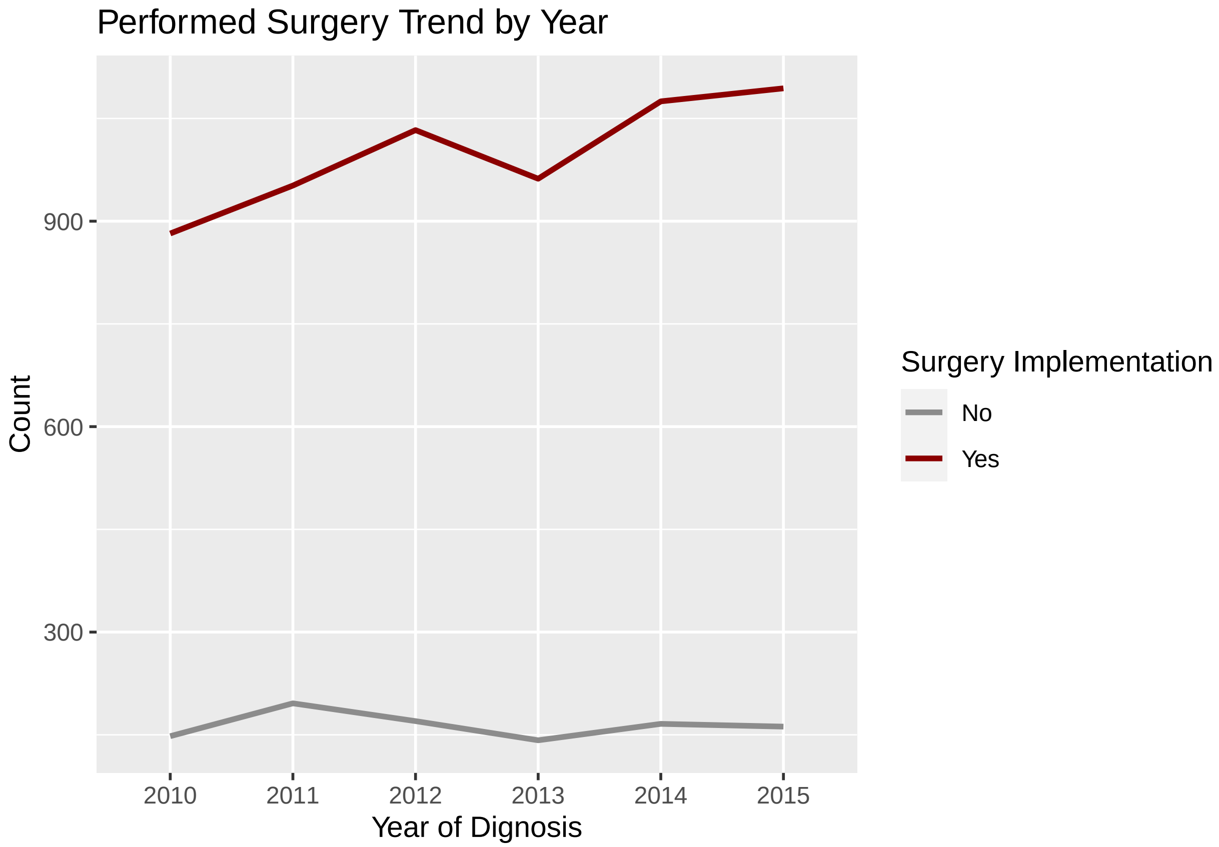


Age and Year of Diagnosis

The age feature is recorded as a continuous variable, so we make a density plot instead of the bar plot. In the below plot, we find that the distributions of performing surgery and not performing surgery on ages are mostly overlapping. However, we can still see that, for people older than 75 years old, they are more likely to perform the surgery; while for people younger than 75 years old, they are more likely to not perform the surgery.

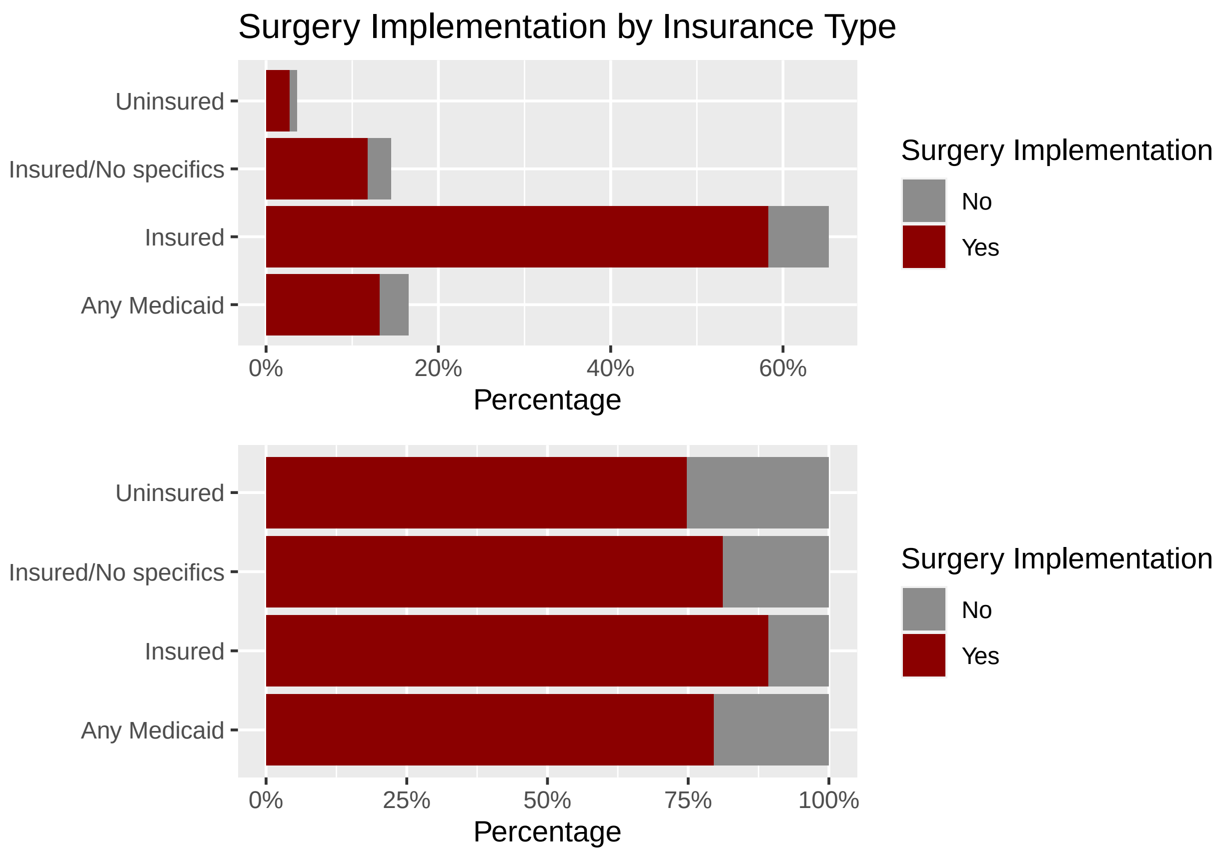


Then, we would like to explore the change of surgery implementation on time series. From the plot below, for the trend of amount of performed surgery, we can see that as time goes by, there is an upper trend.



Insurance

Besides of the demographic features, we also concern about the impact of social biases like insurance on whether to perform the surgery or not. From below chart, we can see that the large majority of repondents are insured and insured people are most likely to perform the surgery.



Feature Engineering

Feature Establishment

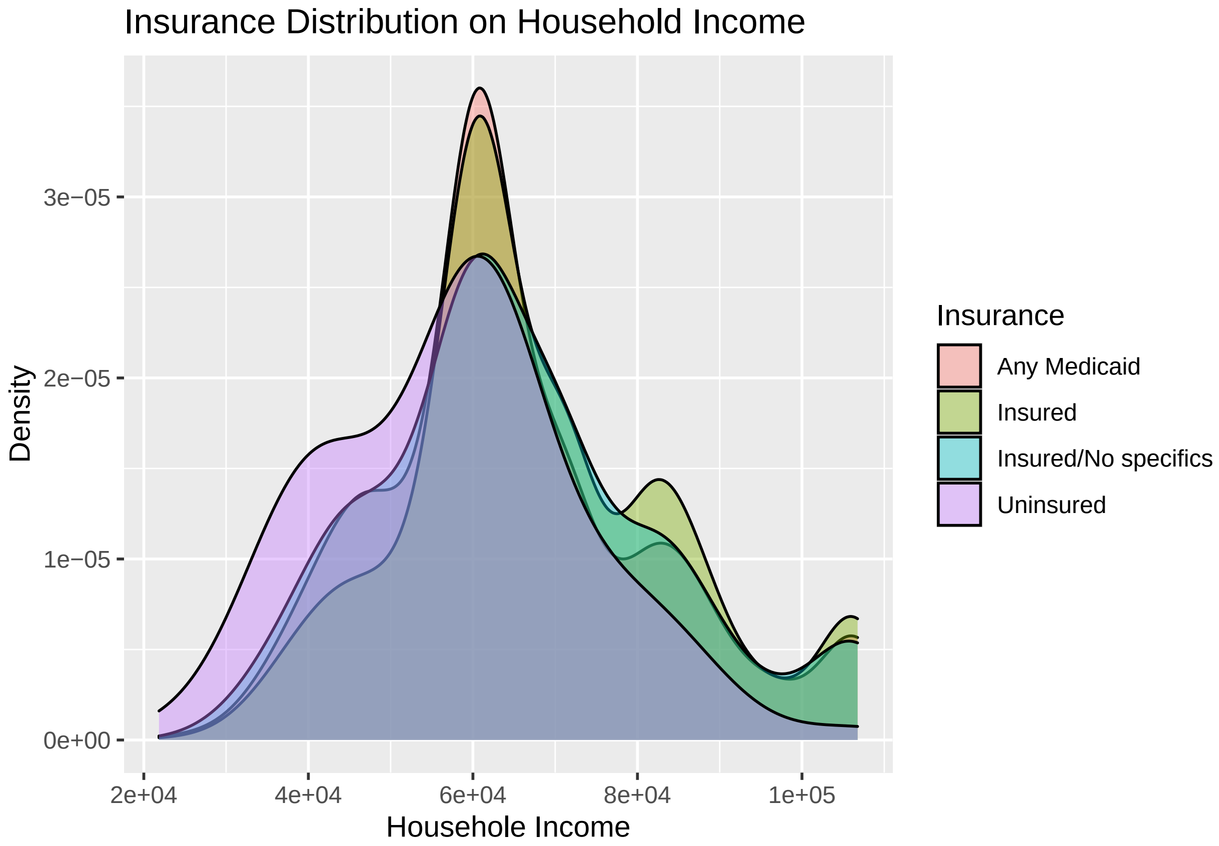
The columns `Surgery.Performed.`,`Radiation` and `Chemotherapy` contain information of the actual therapy that the patients have taken. It would be more informative to combine these 3 columns together to reflect patients' treatment as a whole. Hence, we create a column `True_Therapy` to get the aggregated information.

Meanwhile, the 2021 version **NCCN Guideline** has also provided the information of what kinds of treatments patients with oral cavity cancer are supposed to take. Since the recommended treatment is based on `AJCC.7.Stage`, where contains information of tumor size (`Size`) and lymph nodes status (`Lymph.Nodes`), we will create a column `Rec_Therapy` indicating the treatments that a patient is supposed to have under **NCCN Guideline**.

For equity issue, we believe it would be more important to examine the matchness of recommended therapies and the real therapies that patients actually get. So, we create a column `Match` to evaluate such matchness. `Match` is a binary feature, it equals to 1 when `True_Therapy` and `Rec_Therapy` are completely matched or `Rec_Therapy` is the subset of `True_Therapy`(eg. A patient is supposed to take surgery based on NCCN Guideline, and this patient actully has taken surgery and chemo), and it equals to 0 otherwise.

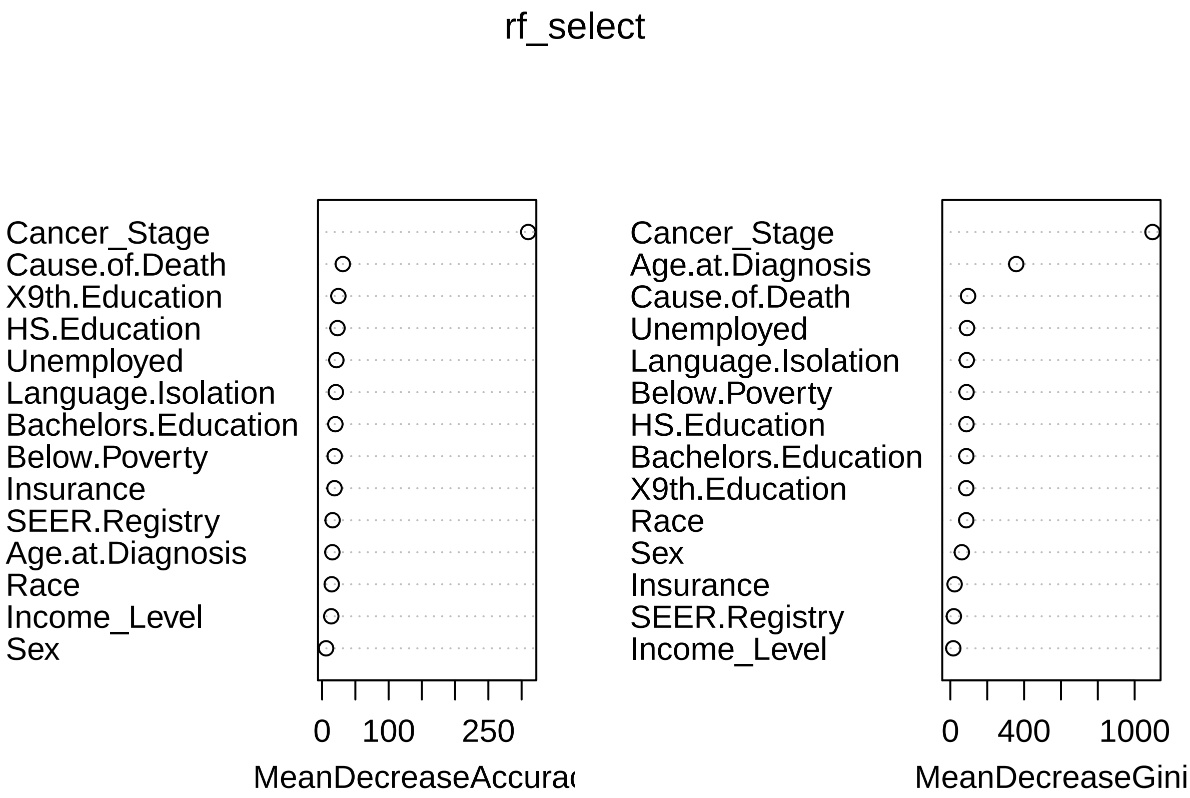
We also do some feature transformations. For the `Median.Household.Income`, we change it from continuous to ordinal with 3 levels. The first level is under first quantile of the income, the second level is between first quanntile and third quantile of the income, and the third level is greater than third quantile of the highest income. For the `Race`, we use the white as the reference group.

Referring to the below plot, different insured conditions have overlapping distributions on the income population from low to high. So we also change the `Insurance` to be binary with 1 representing the patient who had all kinds of insurance and 0 representing the patient who didn't have any insurance. We also change the Sex variable to be binary with 1 representing male and 0 representing female.



Feature selection

When we choose the predictors that will be used in the model, we make a random forest-based selection to see the importance level of all variables. We decide not to use the Income_Level since its importance level is relatively low.



Model

Logistic Regression

Since the outcome is binary, firstly we would like to use the logistic regression model. Except for all the predictors that we decide to use, we also add the interaction term of `Race` and `Sex`.

```
##
## Call:
## glm(formula = Match ~ Cancer_Stage + Language.Isolation + X9th.Education +
##     Bachelors.Education + Unemployed + Age.at.Diagnosis + Below.Poverty +
##     Race + Sex + Race:Sex + Insurance, family = binomial(link = "logit"),
##     data = trainset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2085  -0.5782  -0.3781   0.6898   2.6192
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.551067   0.463549  -1.189  0.23452
## Cancer_StageOral Cavity-II            1.469965   0.105471  13.937  < 2e-16 ***
## Cancer_StageOral Cavity-III           3.839796   0.122130  31.440  < 2e-16 ***
## Cancer_StageOral Cavity-IVA           3.555869   0.099432  35.762  < 2e-16 ***
## Cancer_StageOral Cavity-IVB           2.677457   0.190496  14.055  < 2e-16 ***
## Cancer_StageOral Cavity-IVC           1.464121   0.242658   6.034 1.60e-09 ***
## Cancer_StageOral Cavity-IVNOS          1.806551   0.358426   5.040 4.65e-07 ***
## Language.Isolation                     0.003538   0.024995   0.142  0.88742    
## X9th.Education                        -0.025723   0.028146  -0.914  0.36077    
## Bachelors.Education                   -0.007626   0.006961  -1.096  0.27329    
## Unemployed                            -0.008275   0.027486  -0.301  0.76338    
## Age.at.Diagnosis                      -0.024195   0.002570  -9.414  < 2e-16 ***
## Below.Poverty                         -0.012761   0.011773  -1.084  0.27842    
## RaceAsian or Pacific Islander          0.088143   0.194500   0.453  0.65042    
## RaceBlack                             -0.530187   0.209907  -2.526  0.01154 *  
## RaceHispanic                          -0.155889   0.174066  -0.896  0.37048    
## RaceAmerican Indian/Alaska Native     -0.232158   0.595413  -0.390  0.69660    
## Sex                                   -0.176231   0.089030  -1.979  0.04776 *  
## Insurance                              0.520147   0.178509   2.914  0.00357 ** 
## RaceAsian or Pacific Islander:Sex      0.287099   0.255635   1.123  0.26140    
## RaceBlack:Sex                          0.049180   0.263059   0.187  0.85170    
## RaceHispanic:Sex                       0.537039   0.220622   2.434  0.01492 *  
## RaceAmerican Indian/Alaska Native:Sex  0.076711   0.826572   0.093  0.92606    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 7528.7  on 5584  degrees of freedom
## Residual deviance: 4997.3  on 5562  degrees of freedom
## AIC: 5043.3
## 
## Number of Fisher Scoring iterations: 5
```

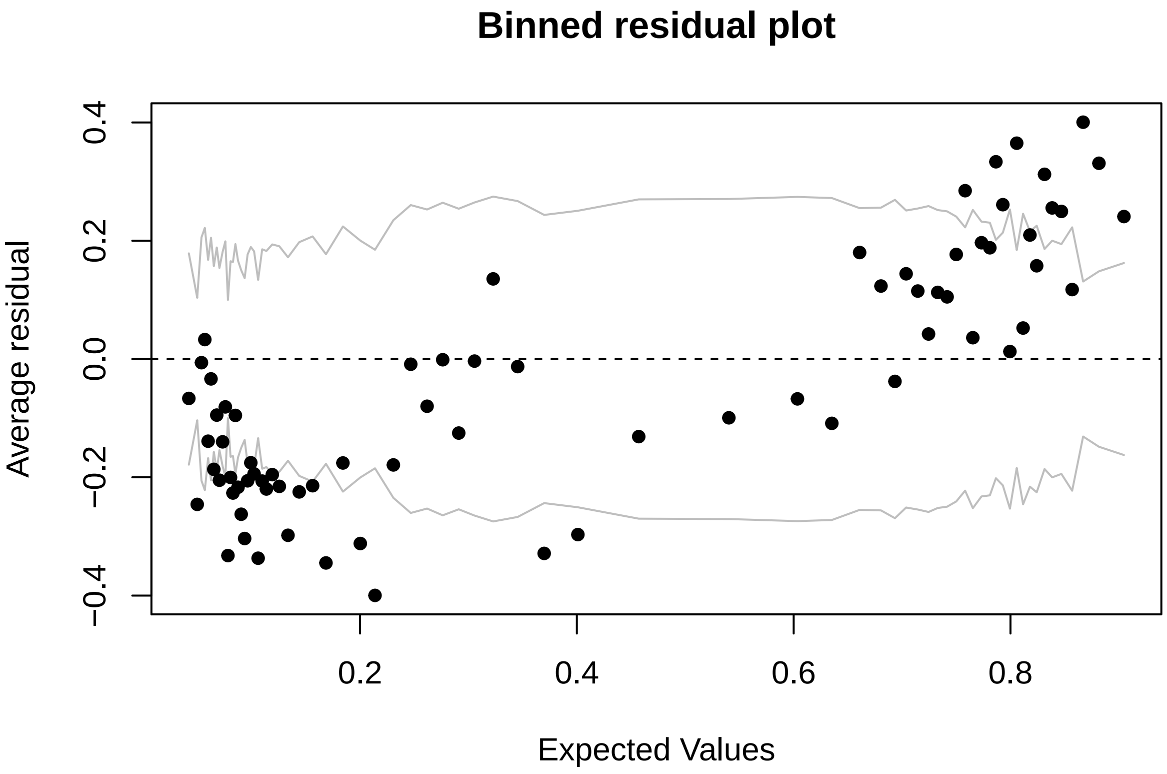
```
##                           (Intercept)            Cancer_StageOral Cavity-II 
##                             0.3656169                             0.8130520 
##           Cancer_StageOral Cavity-III           Cancer_StageOral Cavity-IVA 
##                             0.9789544                             0.9722363 
##           Cancer_StageOral Cavity-IVB           Cancer_StageOral Cavity-IVC 
##                             0.9356833                             0.8121621 
##         Cancer_StageOral Cavity-IVNOS                    Language.Isolation 
##                             0.8589445                             0.5008846 
##                        X9th.Education                   Bachelors.Education 
##                             0.4935697                             0.4980936 
##                            Unemployed                      Age.at.Diagnosis 
##                             0.4979313                             0.4939515 
##                         Below.Poverty         RaceAsian or Pacific Islander 
##                             0.4968099                             0.5220216 
##                             RaceBlack                          RaceHispanic 
##                             0.3704732                             0.4611066 
##     RaceAmerican Indian/Alaska Native                                   Sex 
##                             0.4422198                             0.4560558 
##                             Insurance     RaceAsian or Pacific Islander:Sex 
##                             0.6271820                             0.5712858 
##                         RaceBlack:Sex                      RaceHispanic:Sex 
##                             0.5122925                             0.6311234 
## RaceAmerican Indian/Alaska Native:Sex 
##                             0.5191685
```

Interpretation

From the summary of the logistic regression, we can see that all of the estimated coefficients for `Cancer_Stage` are significant. Especially for stage III, patients are more likely to follow the guideline of therapy compared with other stages. The estimated coefficient for `Age.at.Diagnosis`, `RaceBlack`, `Sex` and `Insurance` are all significant, and the first three coefficients are negative which means that the older patients, blacks and males tend to not follow the guideline of therapy. The estimated coefficient for `Insurance` is positive and significant which means that patients who have insurance are more likely to follow the guideline of therapy than those who don't have insurance. From the prospective of porbability, the blacks are 37% less likely to get correct therapy than that of whites. The patients with no insurance are 63% less likely to follow the guideline of therapy than patients who have insurance. Females are 18% more likely to follow the guideline of therapy than males.

Binned Residual plot

From the binned residual plot, we can see that lots of points are outside the bin which needs further adjustment.



Confusion Matrix

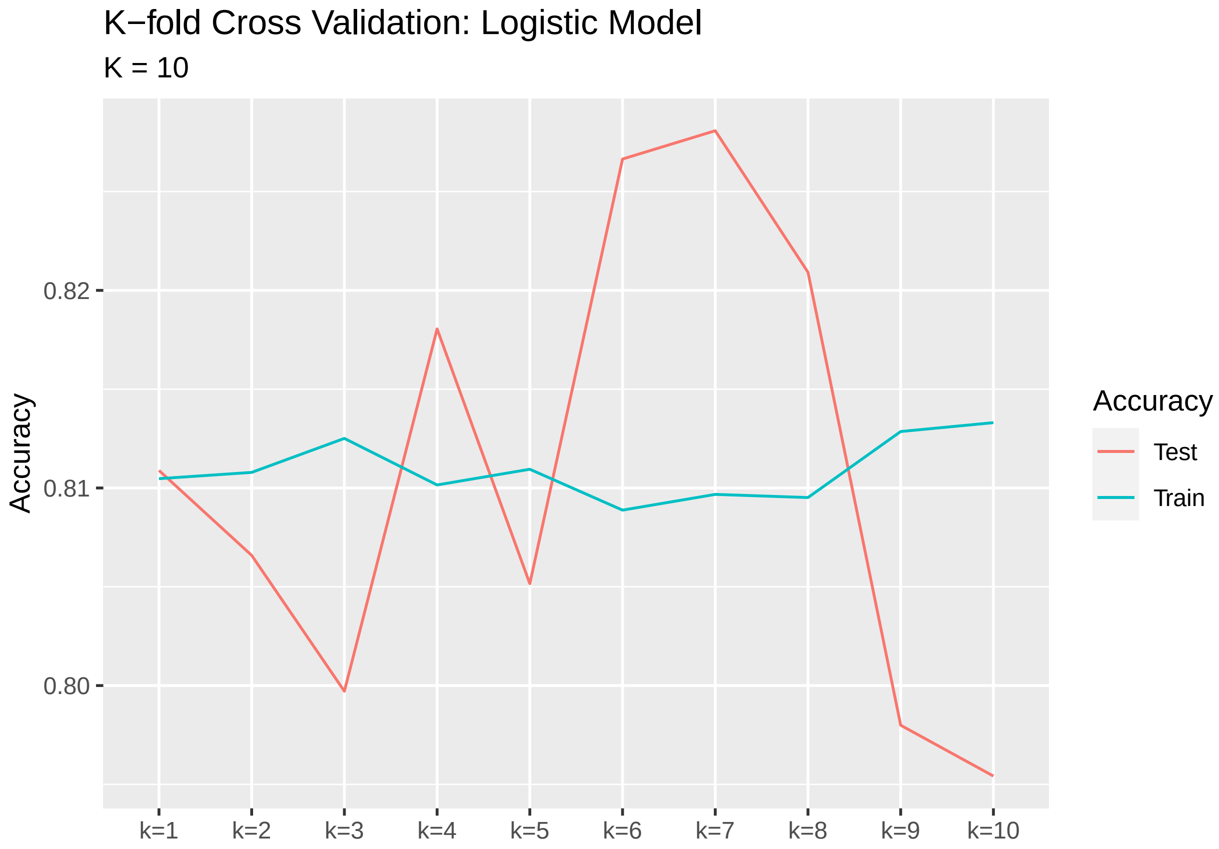
According to the confusion matrix about the test dataset, we see that the prediction accuracy rate is about 80.17% which is pretty high.

```
##
## logis_test_pred   0   1
##               0 690 141
##               1 136 430
```

```
## [1] 0.801718
```

Cross-Validation

To test the sensitivity of logistic model, we use the K-fold cross validation to evaluate the change of accuracy as we use different train set and test set. The result shows that, the accuracy will change dramatically if we adjust the dataset to test the logistic model.



Naive Bayes

Naive Bayesian classifier can take the given variables as conditions to calculate the probability that an observation can be classified into a certain category.

In this project, we have a lot of features that relates to the matching between the therapy a patient is supposed to have and actually have. These features can be seen as given conditions, which are also known as prior information. So, Naive Bayesian classifier would be appropriate for this problem.

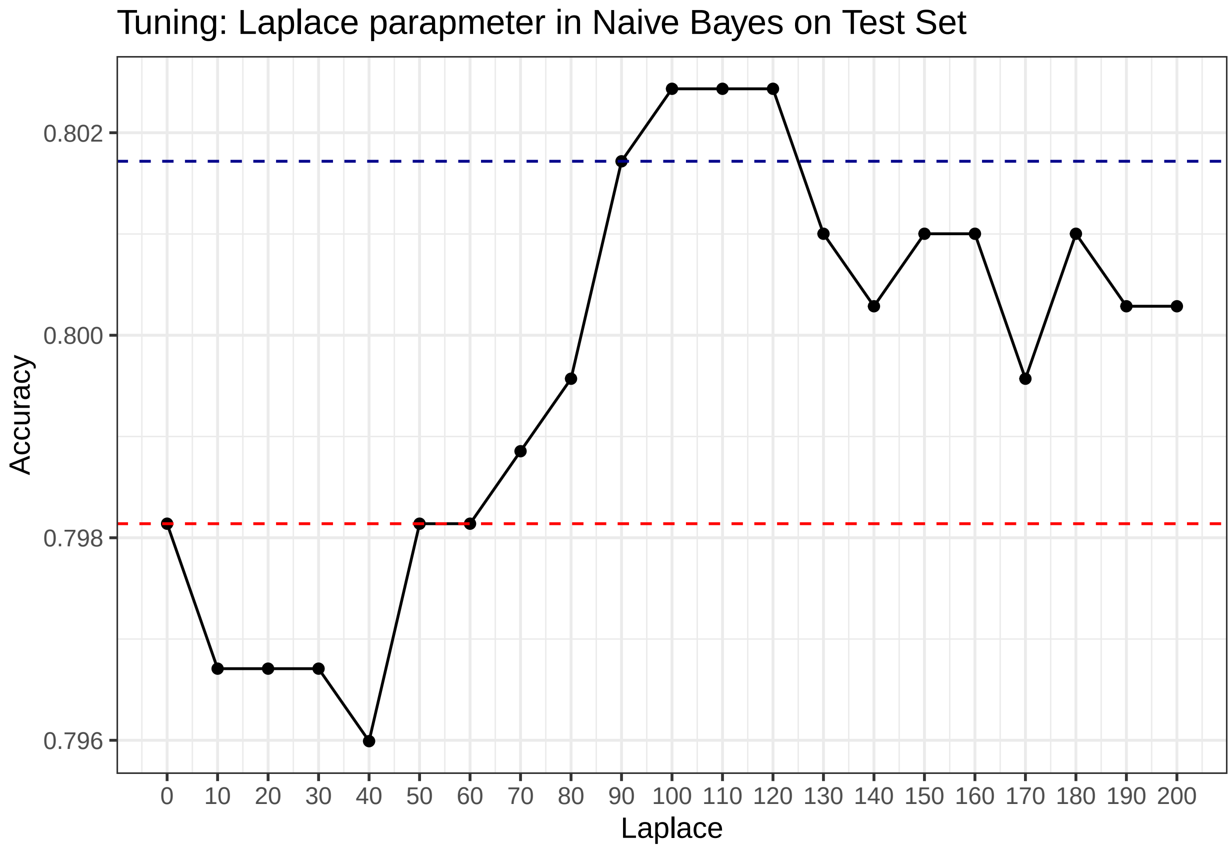
By using `e1071` package, we have a initial Naive Bayesian model with an accuracy of 79.81%, which is slightly lower than the accuracy of the logistic (80.17%).

When looking at the conditional probability in the result of the model, we can see that the probability of matching given $Race = Black$ equals to 8.1%, and probability of matching given $Race = White$ equals to 66.7%, so there is large discrepancy between these two groups. However, white people also have the highest probability of mismatching, which might due to the imbalanced white population in the dataset.

	White	Asian or Pacific Islander	Black	Hispanic	American Indian/Alaska Native
0	0.7398861	0.0761163	0.0650285	0.1126761	0.0062931
1	0.6672598	0.1023132	0.0814057	0.1419039	0.0071174

```
##
## bayes_test_pred   0   1
##               0 690 146
##               1 136 425
## [1] 0.7981389
```

Then we do the tuning of the `laplace` parameter, the red line indicates the accuracy of the initial model, and the blue line indicates the accuracy of the logistic model. When *laplace* = 100, the model reached the highest accuracy which might due to the fact that this model takes conditional probabilities into consideration.



Discussion

Firstly, the dataset itself has contained bias that the majority of the respondents are insured white male from California which result in imbalance. Secondly, the binned residual plot suggests that we still need further research to make the plot look better. Thirdly, about the Naive Bayes, we should have also use k fold cross-validation on the Naive Bayes model, but its computation was extremely time-consuming. Besides, the hypothesis of Naive Bayesian classifier is that all variables' impact on the outcome is independent, but there are interactions among our features. Last but most importantly, there are lots of uncertainties in the dataset which are usually represented by "unknown". For example, in the `Radiation` columns, for many respondents, whether the radiation conducted before or after the surgery is unknown. Regardless of these uncertainties, we can research whether the order between radiation and surgery has impact on the therapy or not.

Conclusion

In general, we can conclude that there exists bias in this data since we know that blacks are less likely to get incorrect therapy compared with other races from the result of the model. We can also see that patients with no insurance are less likely to take treatments. Besides, there are some subtle trends that patients with low educational level and patients who are unemployed are slightly tend to not follow the guideline of therapy.

Reference

(1) Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

(2) Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

(3) Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr

(4) Simon Garnier (2018). viridis: Default Color Maps from 'matplotlib'. R package version 0.5.1. https://CRAN.R-project.org/package=viridis

(5) Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3,.

(6) Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 https://mc-stan.org/rstanarm.

(7) Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J & Wolfe R. Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA. https://github.com/stan-dev/stancon_talks/

(8) Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

(9) A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

(10) Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

(11) David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2020). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4. https://CRAN.R-project.org/package=e1071

(12) National Comprehensive Cancer Network. (2021). Available from https://www.nccn.org/