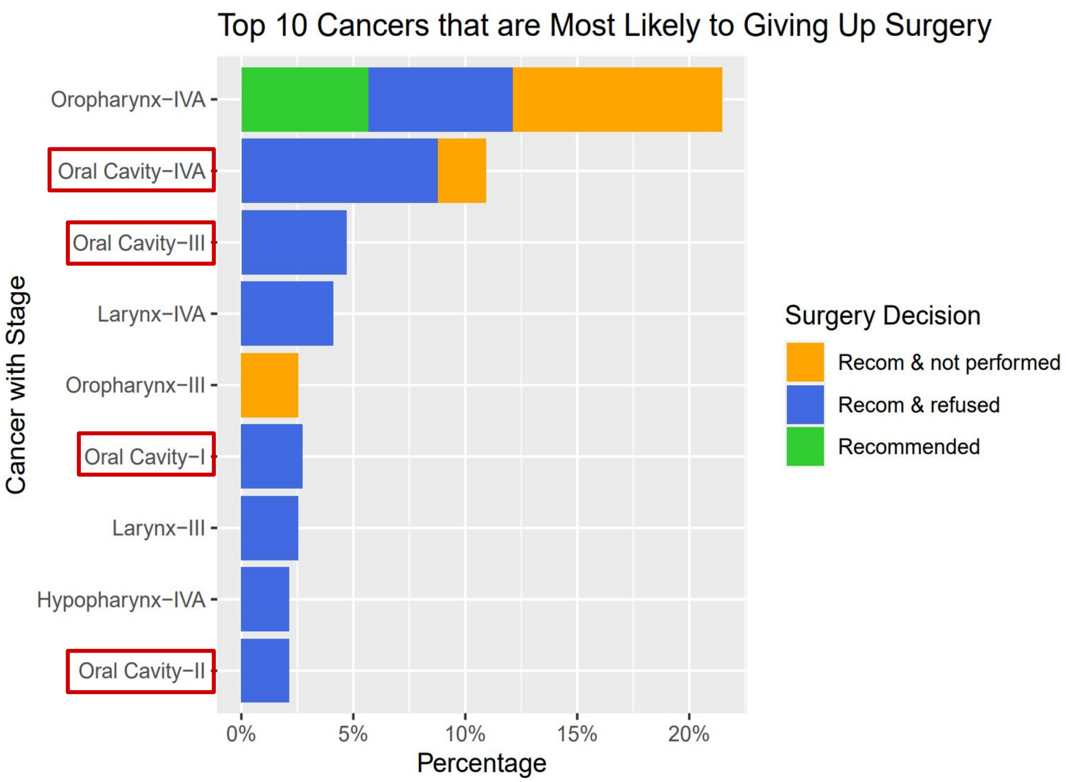
Seer Project

Team 9

Team goal

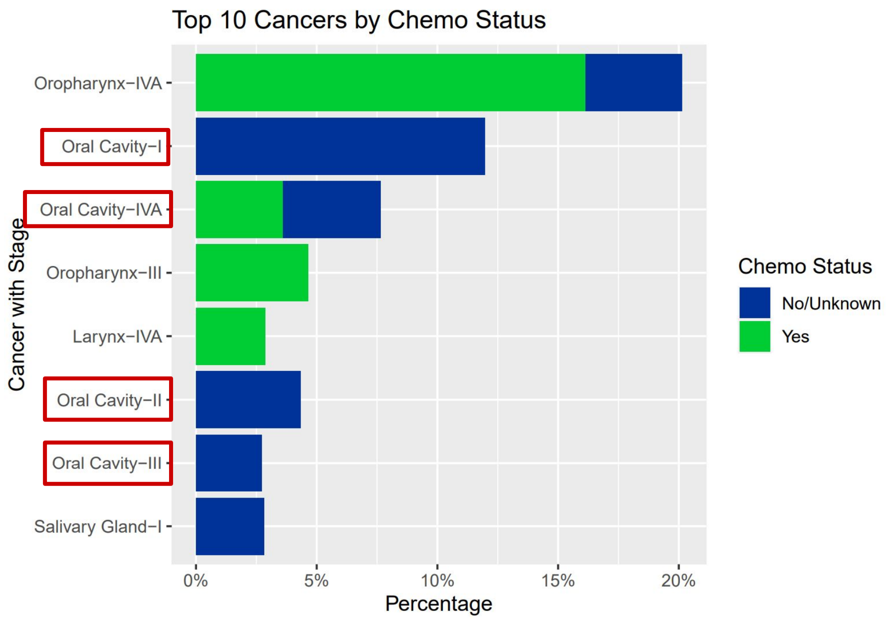
Our main goal of the project is to explore how biases (race, gender, education level and other factors including) will affect the matching of recommended and actual therapy of patients.

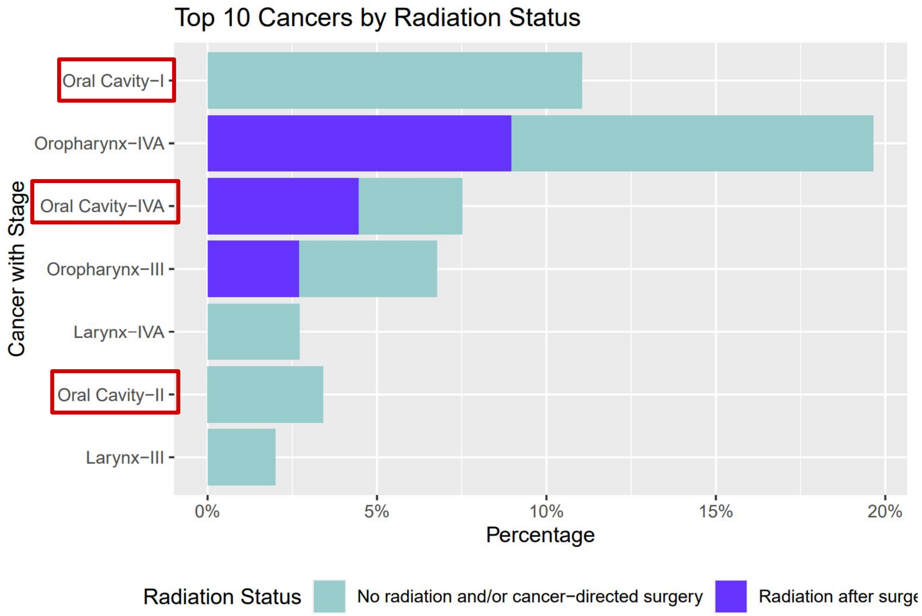
Exploratory Data Analysis



- to simplify our research, we would like to choose only one kind of cancer to analyze
- we firstly explore the top 10 cancers that the patients are most likely to give up the surgery

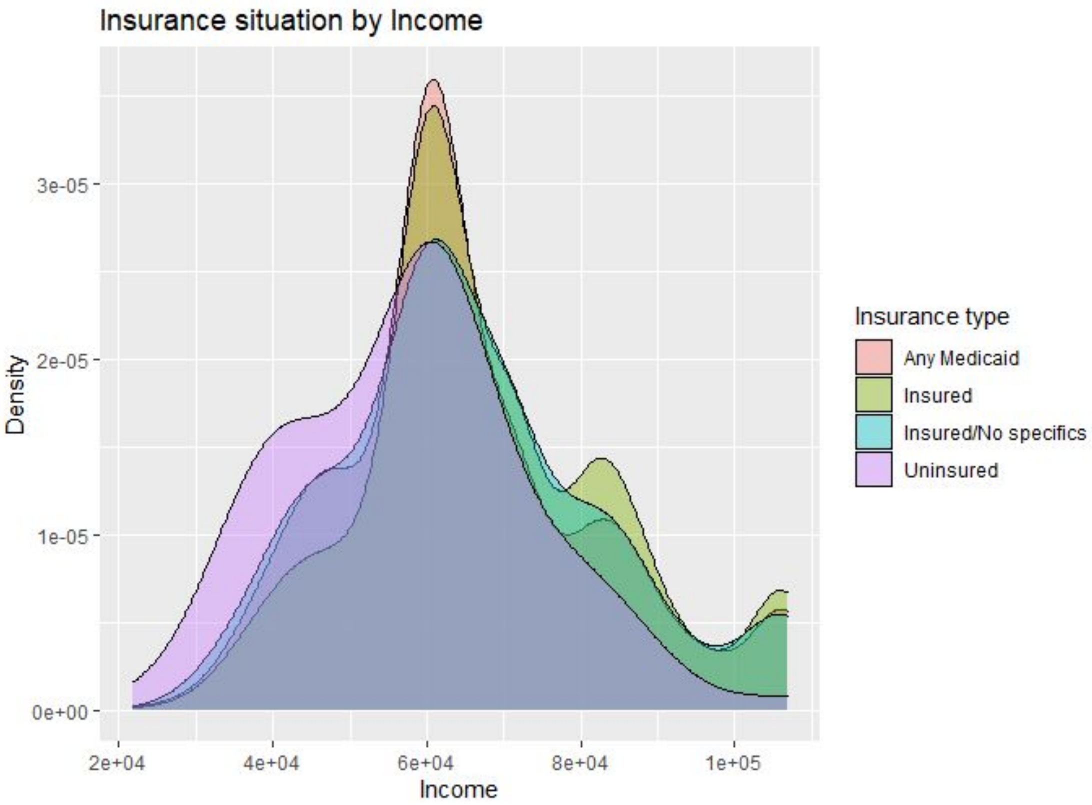
Exploratory Data Analysis





- combining the information given by the status of chemo and radiation therapy, we would like to choose oral cavity cancer as our research direction

Exploratory Data Analysis



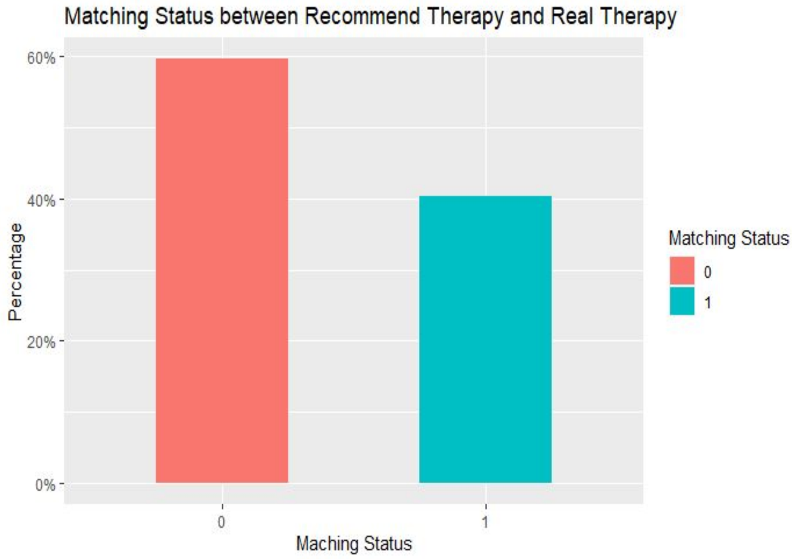
- The overlap of insurance status become more serious as the income increase

Feature Engineering: Establishment

Known information

Created new variables

Known information	Created new variables
Radiation **Chemotherapy** **Surgery.Performed**	**True_Therapy**
what kinds of treatments patients with oral cavity cancer are supposed to take (according to the NCCN Guideline)	**Rec_Therapy**
whether recommended therapies and the real therapies that patients actually get matches	**Match**



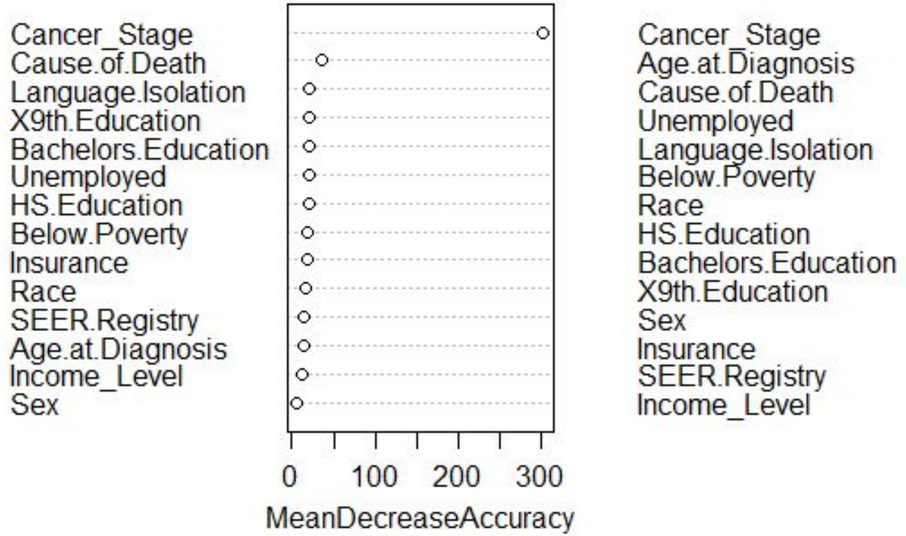
Feature Engineering: Transformation

Median.Household.Income (Numeric Variable) → Income_Level (Ordinal Variable)

Sex-Categorical Variable → Sex-Binary Variable (0 - 1)

Insurance-Categorical Variable → Insurance-Binary Variable (0 - 1)

Feature Selection



- use Random Forest method to select the variables for modeling
- besides, we add Race, Sex and Race:Sex to our model

Model: Logistic Regression

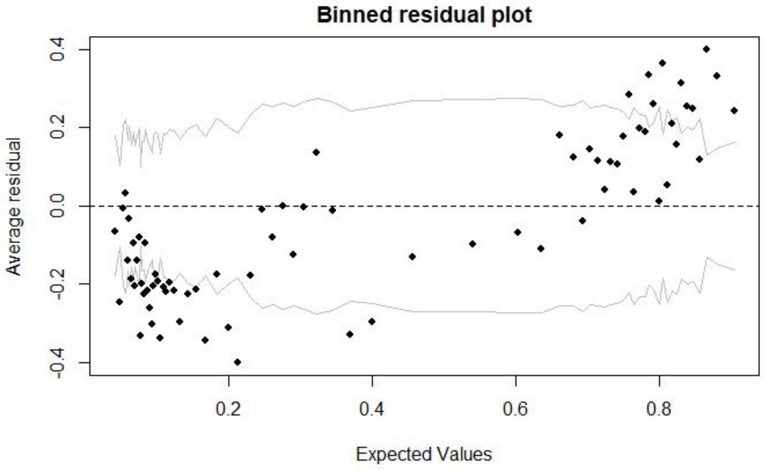
```
m1 <-  glm(Match ~ Cancer_Stage  + Language.Isolation +
X9th.Education + Bachelors.Education + Unemployed + Age.at.Diagnosis
+ Below.Poverty + Race + Sex + Race:Sex + Insurance, data =
trainset, family = binomial(link = "logit"))
```

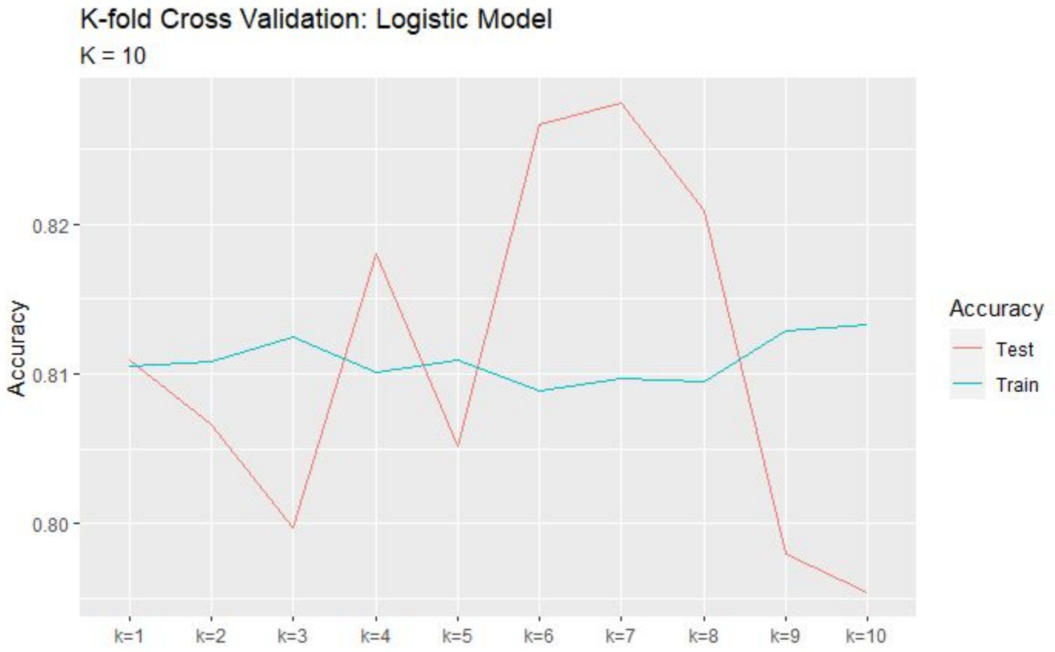
Model: Logistic Regression

```
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    
---
```

- For equity issue, we find that the coefficients of **Age**, **RaceBlack**, **Sex**, **Insurance** and **RaceHispanic:Sex** are significant

Model: Logistic Regression



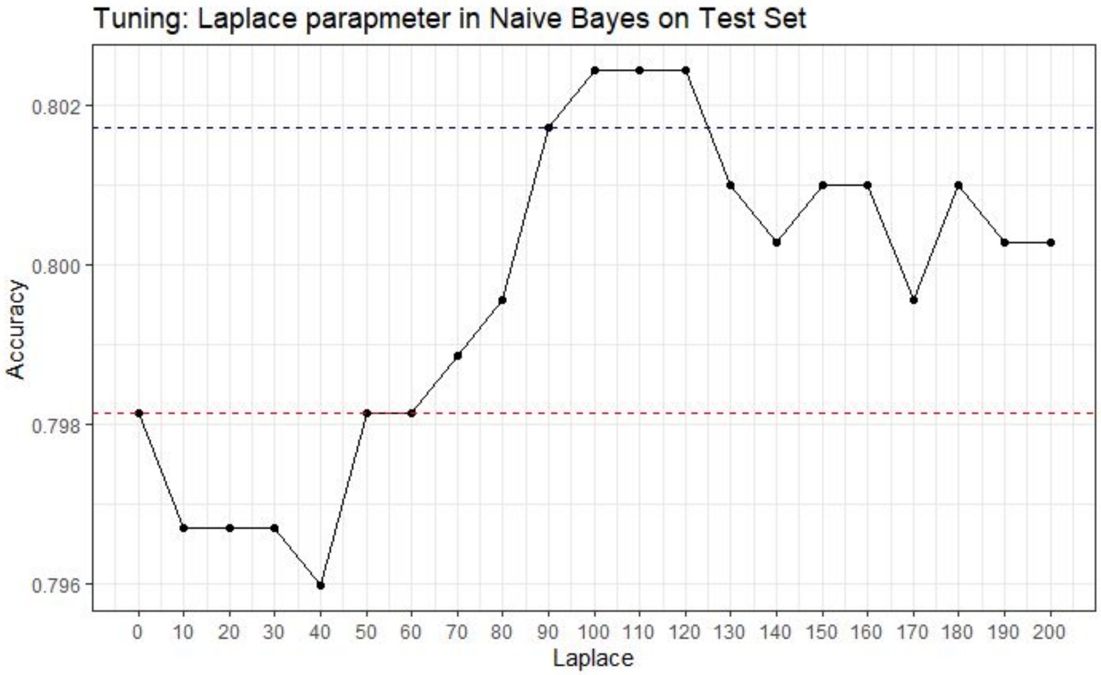


Model: Naive Bayes

```
m2 <- naiveBayes(as.factor(Match) ~ Cancer_Stage  +
Language.Isolation + X9th.Education + Bachelors.Education +
Unemployed + Age.at.Diagnosis + Below.Poverty + Race +
as.factor(Sex) + as.factor(Insurance),data = trainset)
```

Match	White	Asian & PI	Black	Native
0	0.73	0.08	0.07	0.006
1	0.67	0.10	0.08	0.007

Model: Naive Bayes



Blue Line: Logistic Model

Red Line: Naive Bayes Classifier, Untuned

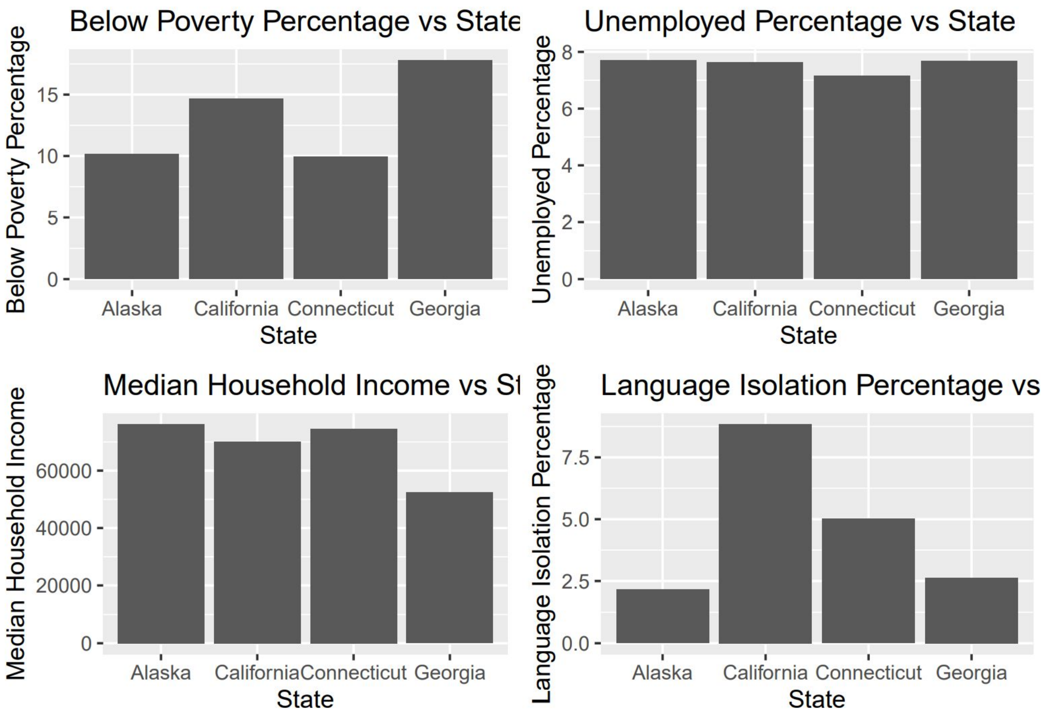
Conclusion

- Bias in this data existed, since we know that blacks are less likely to get correct therapy compared with other races from the result of the model.

- Patients with no insurance are less likely to take treatments.

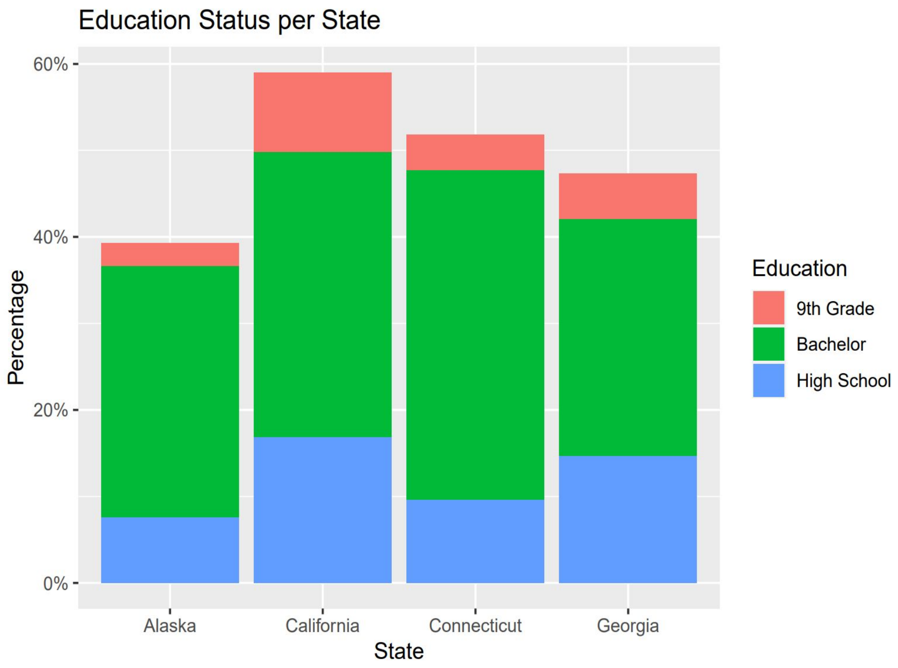
- There are some subtle trends that patients with low educational level and patients who are unemployed are slightly tend to not follow the guideline of therapy.

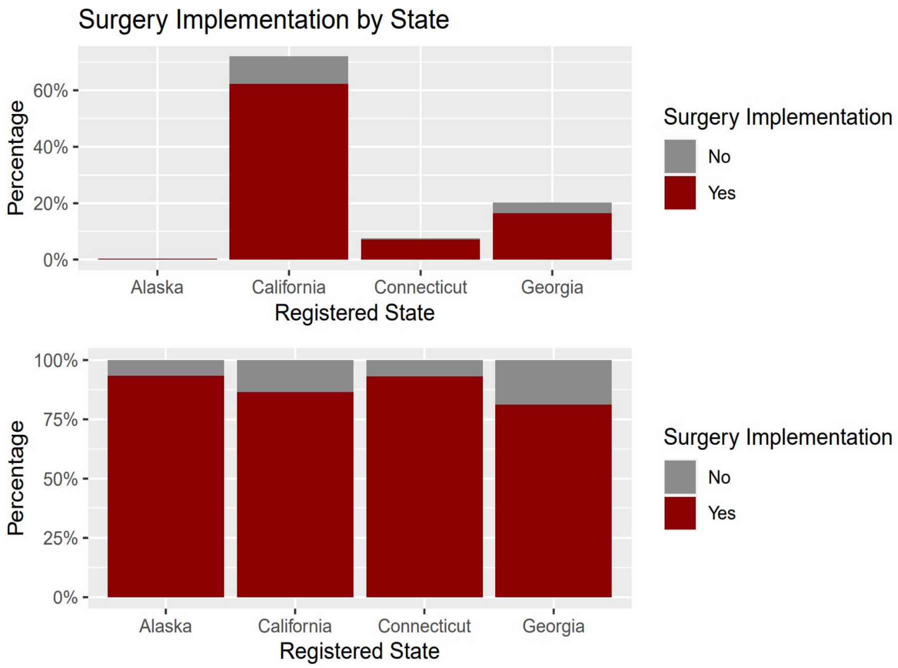
Appendix



- **Environment Attribute vs State :**
  poverty, unemployment, household income and language isolation distribution by states

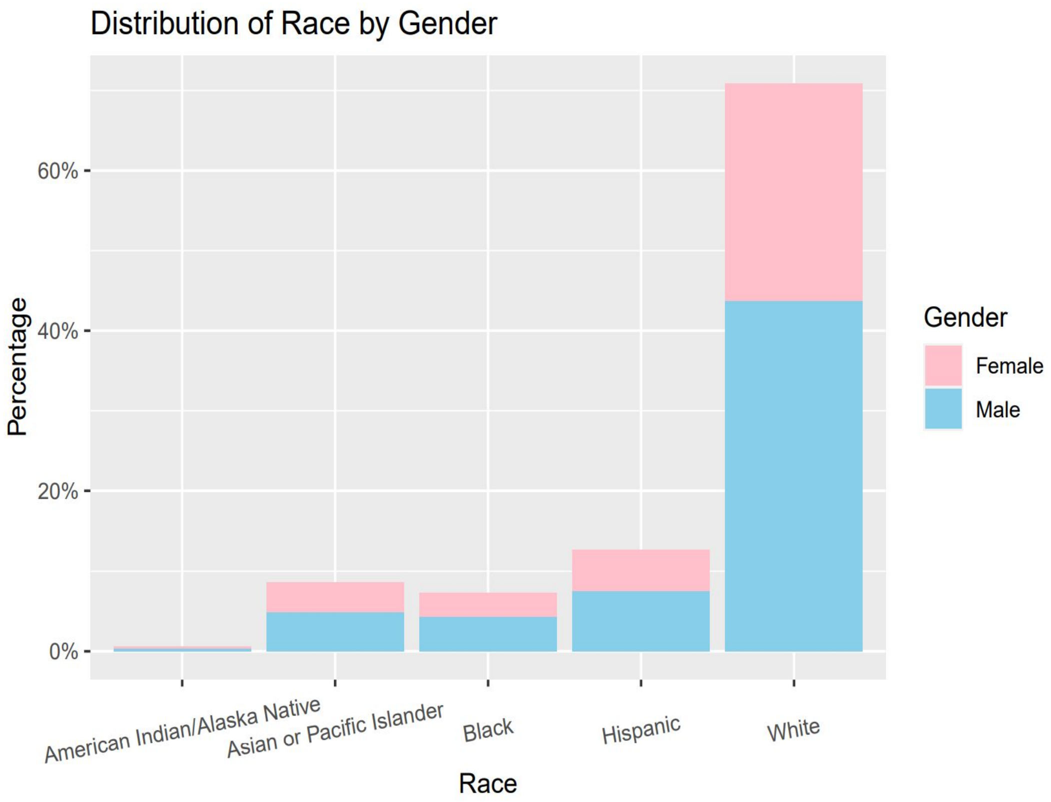
Appendix





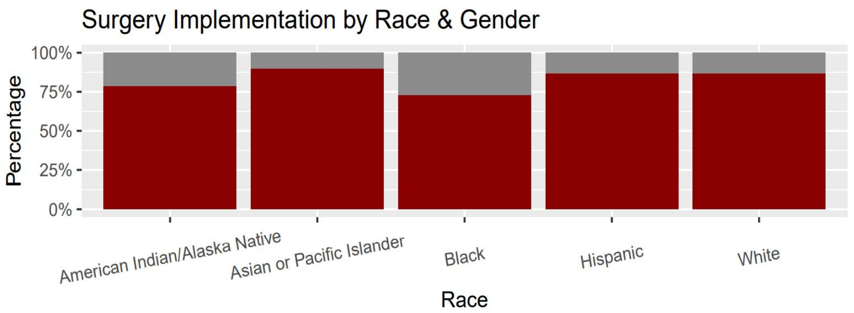
- **Environment Attribute vs State :**
  education status, surgery implementation distribution by state

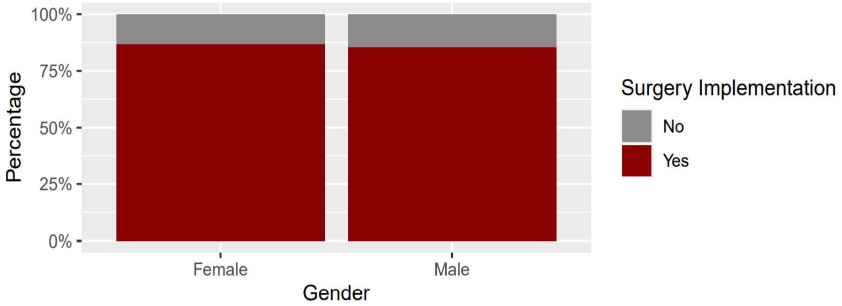
Appendix



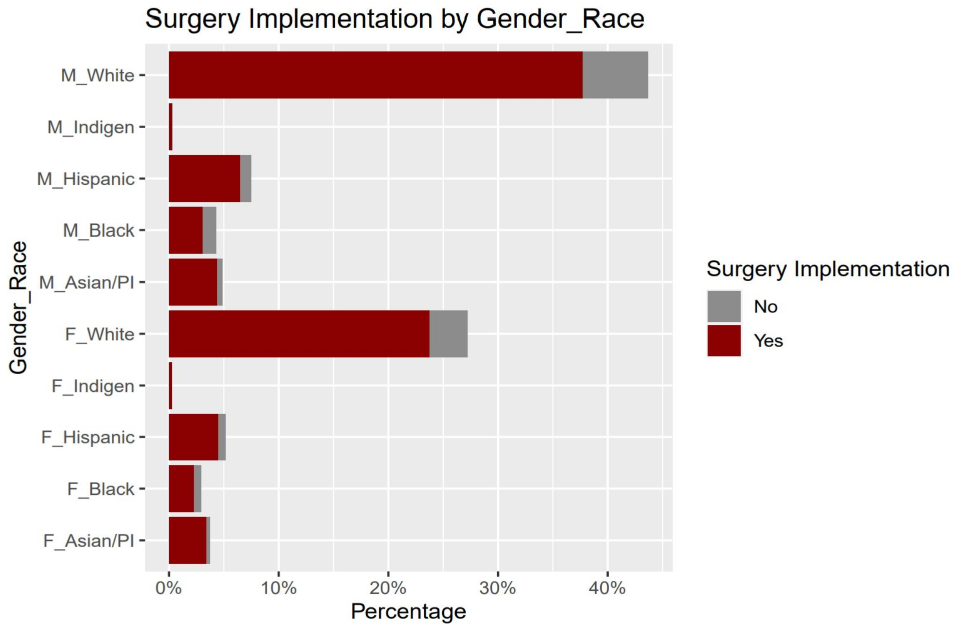
- more than 60% of observations are white people; more than 40% of observations are white male
- there would be an imbalance problem in the data

Appendix



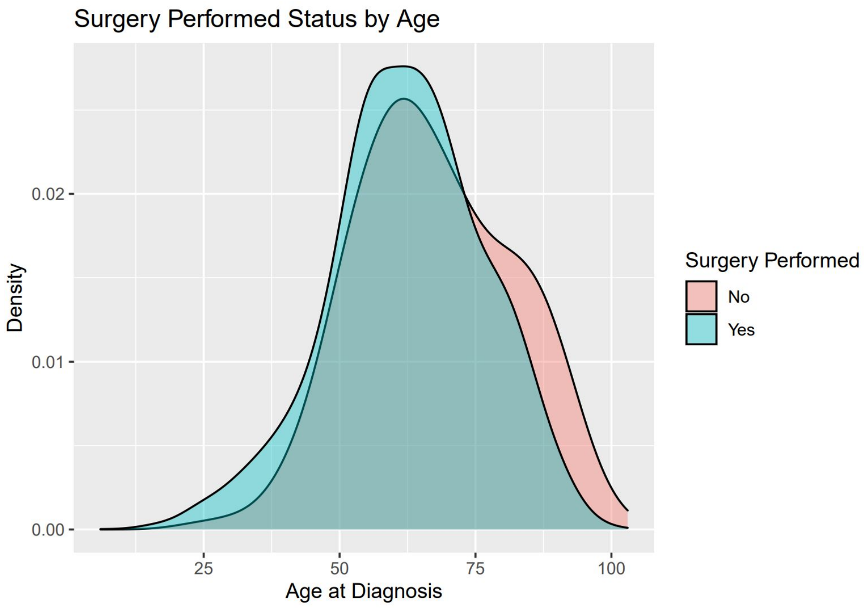


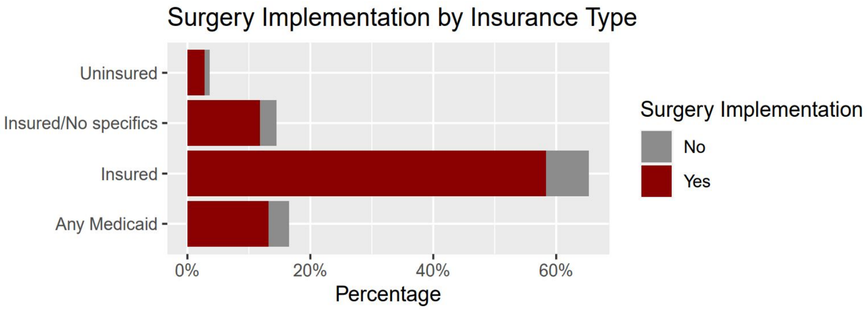
- surgery implementation by race
- surgery implementation by gender

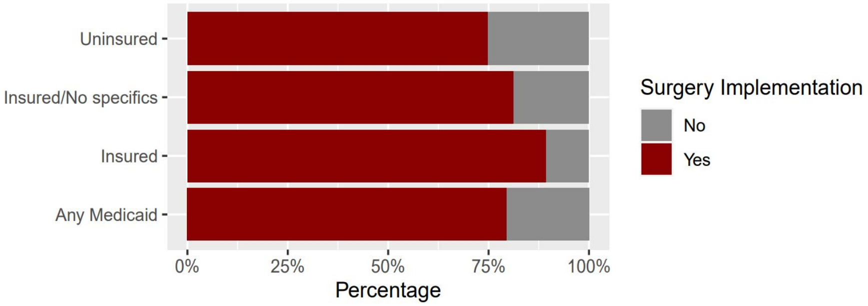


- put these gender and race together to explore the distribution of performed surgery
- white male has the largest proportion of performing a surgery

Appendix



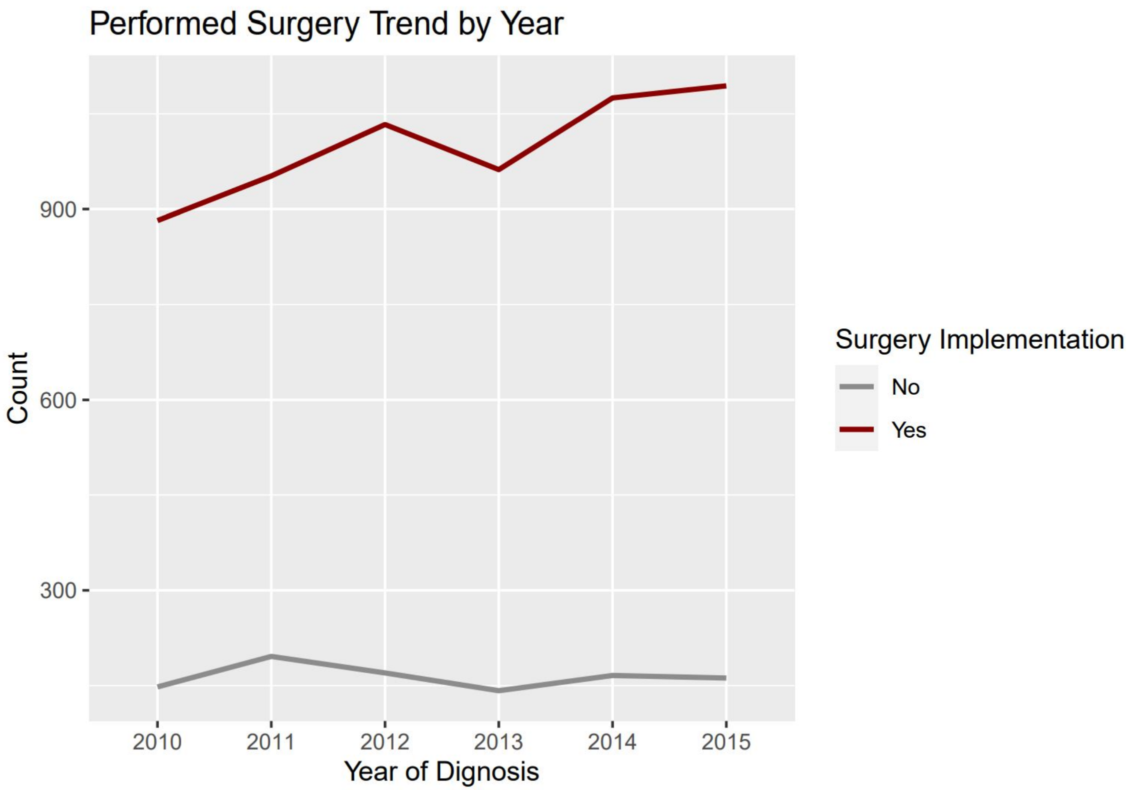




- use density plot to find the distribution of the continuous variable - age
- the distributions of performing surgery and not performing surgery on ages are mostly overlapping

- the large majority of respondents are insured
- insured people are most likely to perform the surgery

Appendix



- The line chart shows the change of surgery implementation over time series
- As time goes by, there is an upper trend for the amount of performed surgery