Predictive models for cumulative COVID-19 cases in USA

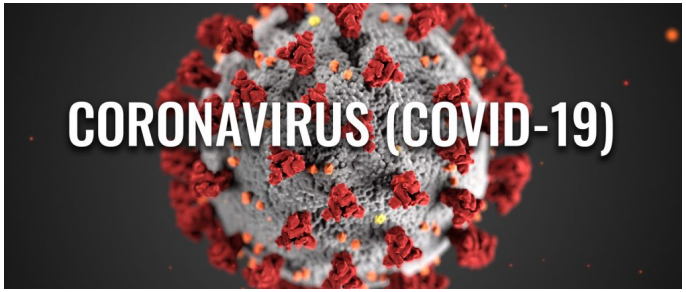
Minsu Kim, Ariane Stark

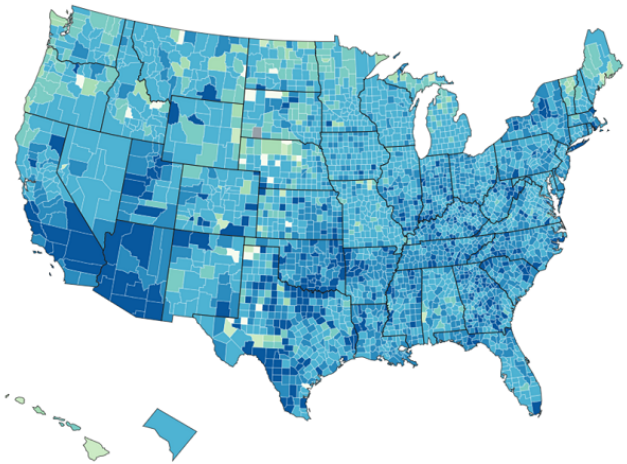
Section 1

Introduction

Background

- COVID-19 is a coronavirus identified in 2019 and has caused a pandemic of respiratory illness.
- It has been spreading rapidly since the first case of COVID-19 was reported Dec. 1, 2019.
- Vaccines have recently been available, but it is too early to be relieved due to concerns about the mutant virus.
- There is still a need to keep an eye on the trends of cases.





(Figure: https://www.denhamblythe.com/covid-19-health-statement) (Figure:https://covid.cdc.gov/covid-data-tracker/#datatracker-home)

Data Set Overview

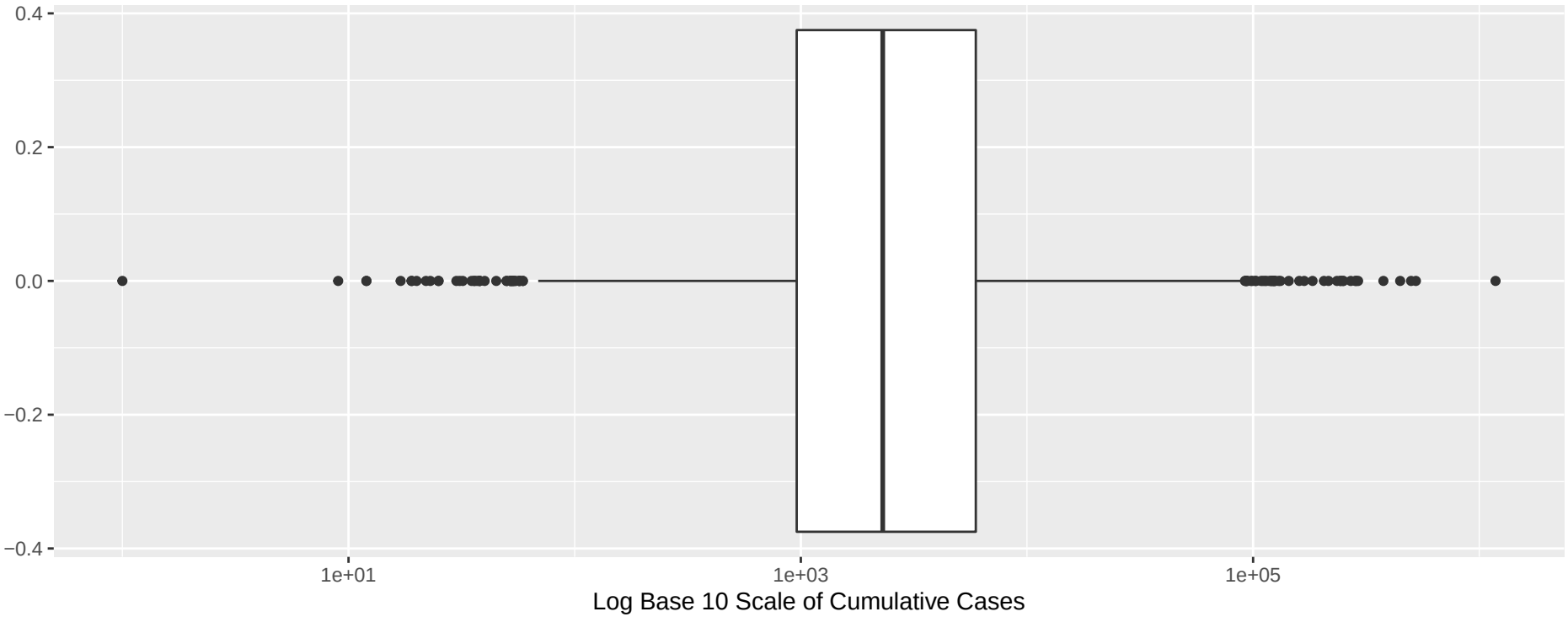
The data set contains cumulative COVID-19 case counts as of April 1, 2021 for each of the 3,220 counties in the United States. Also we have racial and ethnicity demographics, poverty demographics, population demographics, and age demographics from government census data. Note Rio Arriba County, New Mexico is missing poverty demographics and has thus been excluded from analysis, leaving our data set with 9 predictors and 3219 observations.

(Source:https://covid19.census.gov)

Our Variables

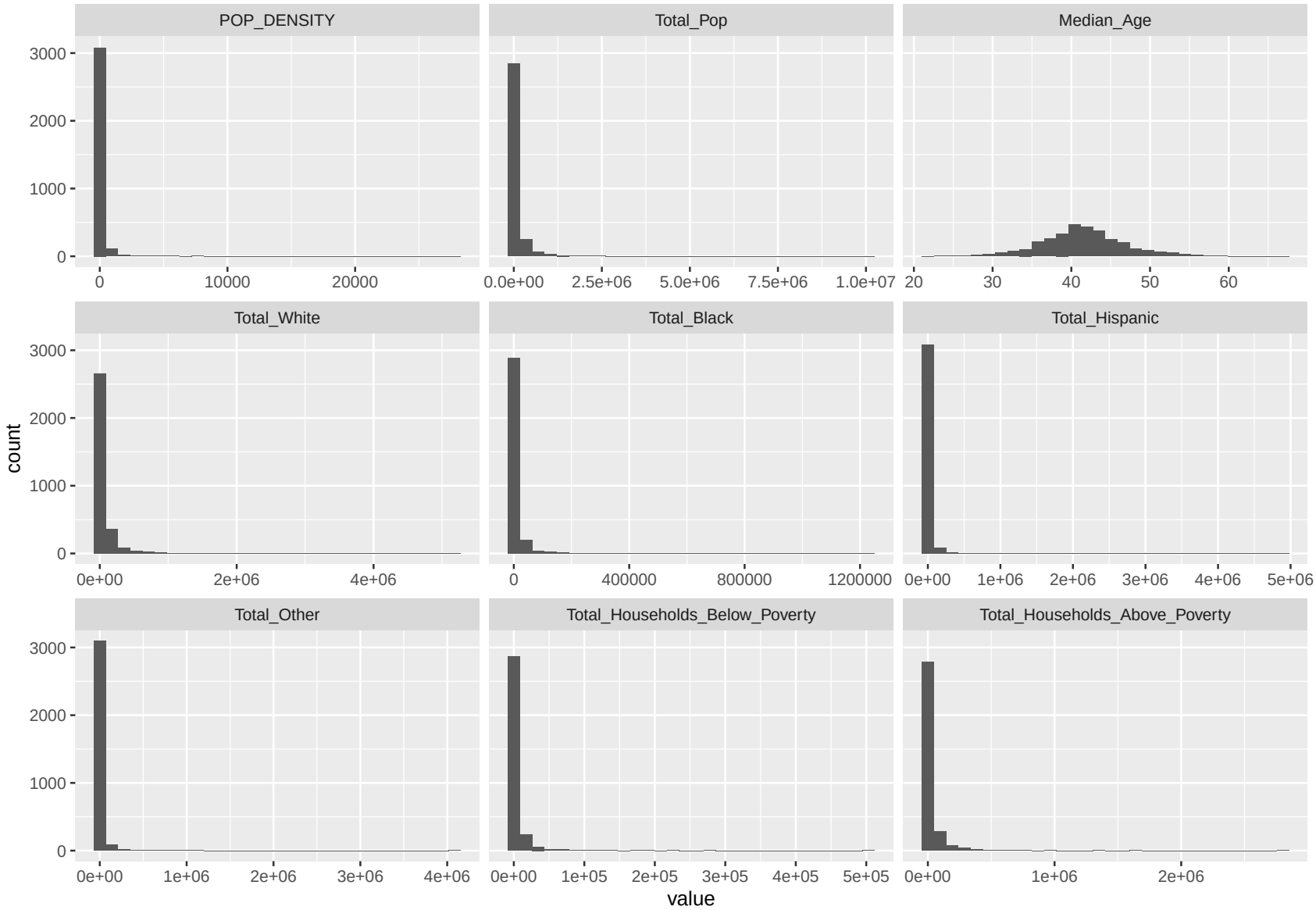
- Cumulative COVID-19 Cases
- Total Population
- Population Density
- Median Age
- Total Number of White Residents
- Total Number of Black Residents
- Total Number of Other (Not strictly White or Black) Residents
- Total Number of Hispanic Residents
- Total Number of Households Above the Poverty Line
- Total Number of Households Below the Poverty Line

Our Variables: Plots (Response)




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0     954    2301    9323    5939 1180538
```

Our Variables: Plots (Predictors)



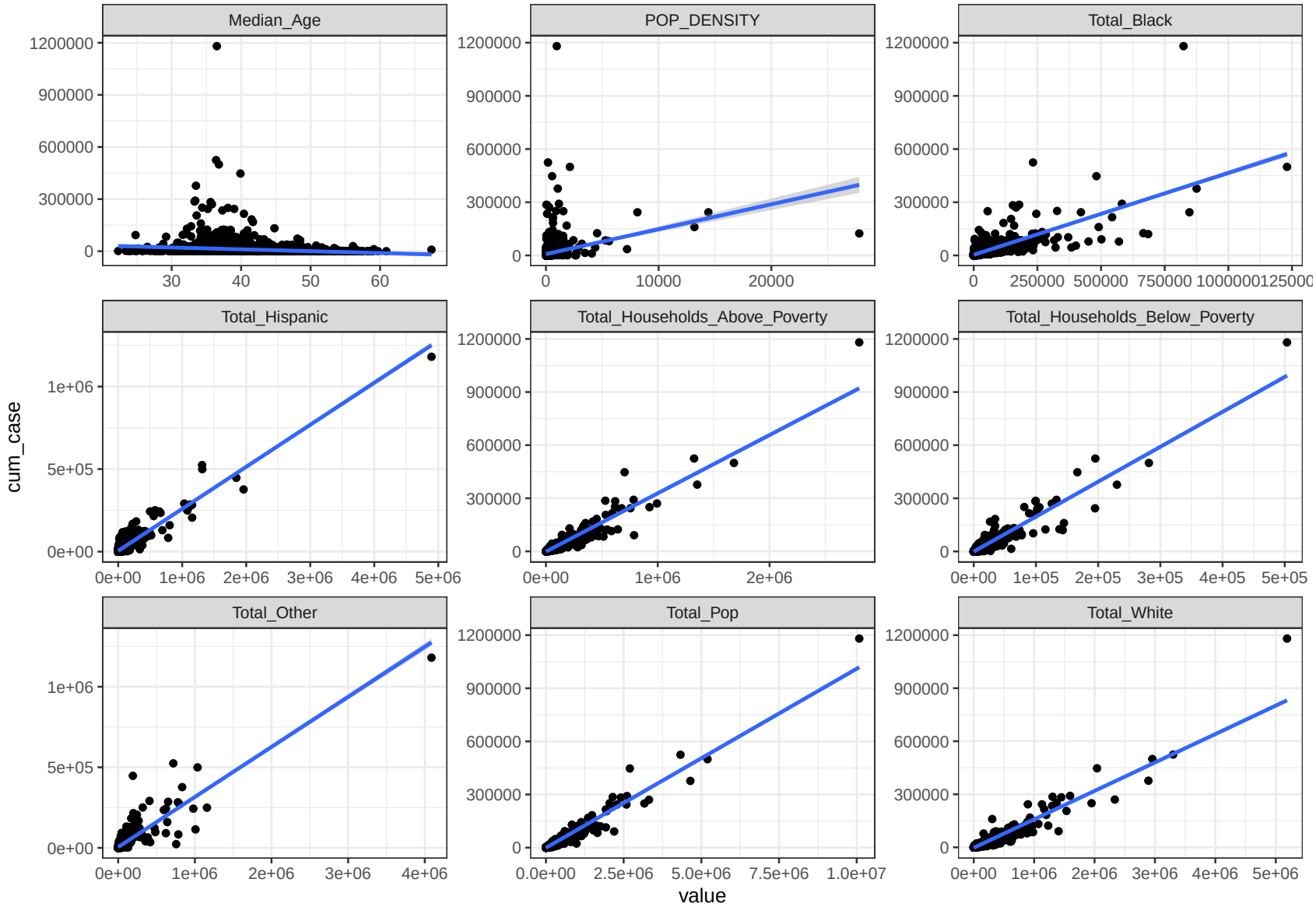
Study Objectives

- Select the demographics that yields a good predictive model for cumulative COVID-19 cases at the county level.
- Fit classification models to predict whether a given county has an instance rate higher than the median of the instance rates.

Section 2

Simple Linear Regression

Simple Linear Regression

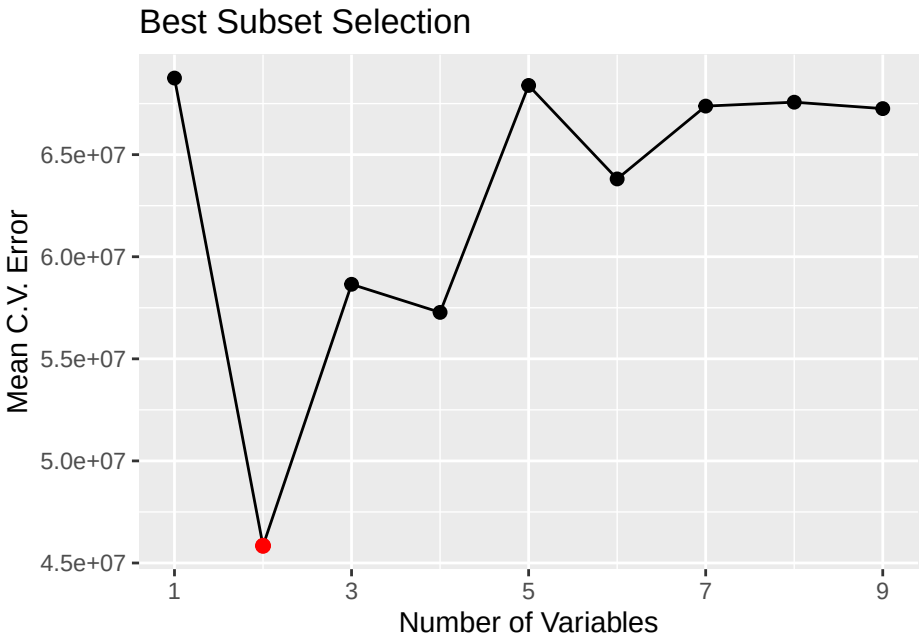


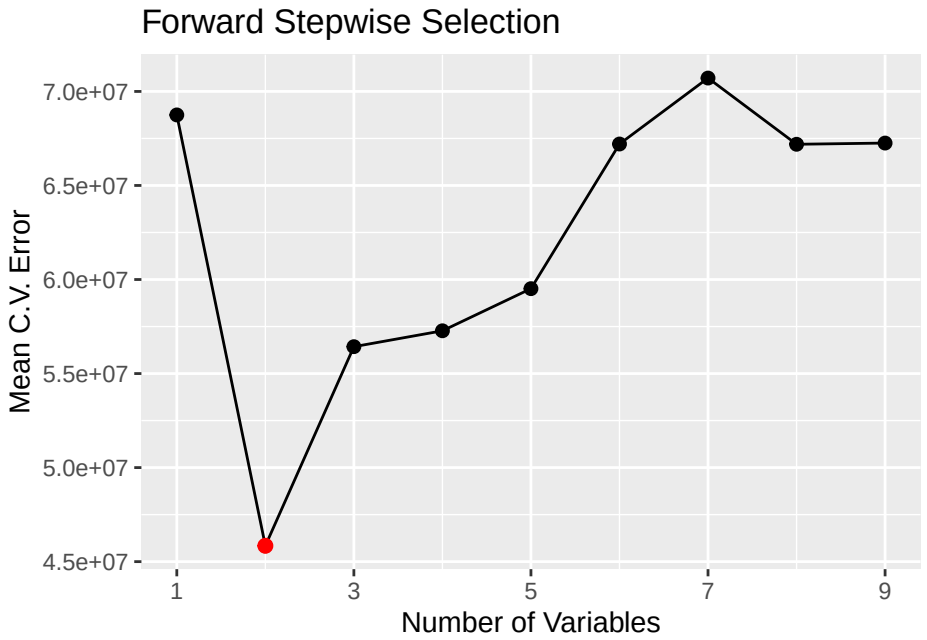
Section 3

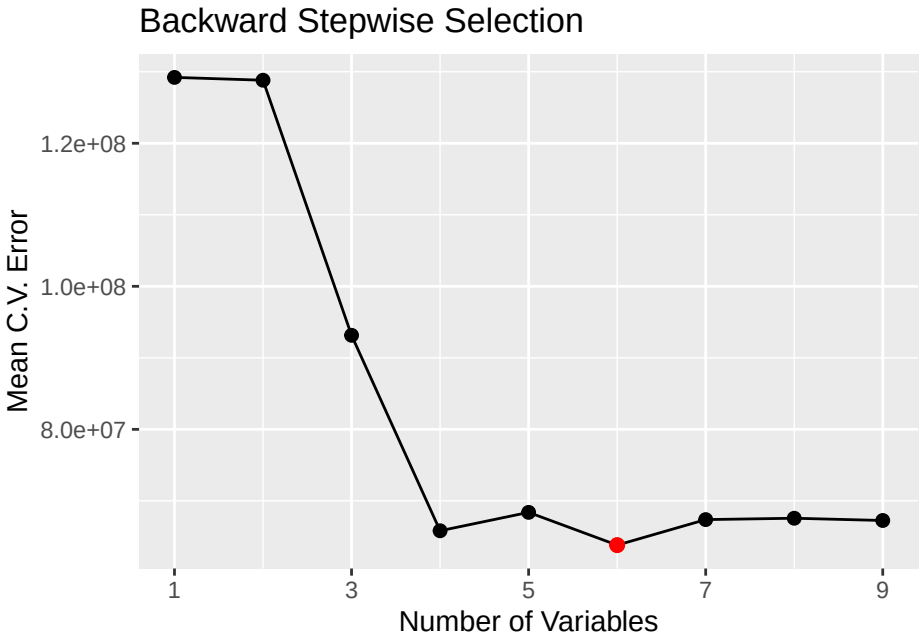
Subset Selection

Cross Validation of Best Subset, Forward and Backward Stepwise

10-fold cross validation of data was done to find the optimal number of parameters for the model selected using each method. This is done to balance the model fitting the data well and the model having decent predictive accuracy.


Best Subset Selection
Mean C.V. Error
6.5e+07
6.0e+07
5.5e+07
5.0e+07
4.5e+07
1
3
5
7
9
Number of Variables


Forward Stepwise Selection
Mean C.V. Error
7.0e+07
6.5e+07
6.0e+07
5.5e+07
5.0e+07
4.5e+07
1
3
5
7
9
Number of Variables


Backward Stepwise Selection
Mean C.V. Error
1.2e+08
1.0e+08
8.0e+07
1
3
5
7
9
Number of Variables

Best Subset Overall

Recall 10 fold cross validation of best subset selection selected the model with 2 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9452	0.9452	2488.4481	-9330.249
2	0.9639	0.9638	546.0689	-10664.172
3	0.9662	0.9662	306.2665	-10870.117
4	0.9674	0.9674	178.4008	-10983.143
5	0.9683	0.9683	89.4799	-11062.648
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Forward Stepwise Overall

Recall 10 fold cross validation of forward stepwise selection selected the model with 2 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9452	0.9452	2488.4481	-9330.249
2	0.9639	0.9638	546.0689	-10664.172
3	0.9662	0.9662	306.2665	-10870.117
4	0.9674	0.9674	178.4008	-10983.143
5	0.9675	0.9674	175.8868	-10979.359
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Backward Stepwise Overall

Recall 10 fold cross validation of backward stepwise selection selected the model with 6 parameters as having the lowest average cross validation predictive error.

Parameters	R^2	AdjR^2	CP	BIC
1	0.9025	0.9025	6923.4012	-7478.489
2	0.9559	0.9559	1375.4685	-10022.384
3	0.9648	0.9647	452.3814	-10739.096
4	0.9665	0.9665	270.9103	-10896.412
5	0.9683	0.9683	89.4799	-11062.648
6	0.9691	0.9691	5.6465	-11139.497
7	0.9691	0.9691	6.1421	-11132.929
8	0.9691	0.9691	8.0746	-11124.920
9	0.9692	0.9691	10.0000	-11116.918

Model Summary

Best Subset: The model has parameters: Total Population, and Total Hispanic.

Cumulative Cases = 315.11931521 + 0.07300349(Total Population) + 0.08304726(Total Hispanic)

Forward Stepwise: The model has parameters: Total Population, Total White, and Total Hispanic.

Cumulative Cases = 315.11931521 + 0.07300349(Total Population) + 0.08304726(Total Hispanic)

Backward Stepwise: The model has parameters: Total Population, Total White, Total Black, Total Hispanic, Total Other, and Total Number of Households Above Poverty.

Cumulative Cases = -338.60412493 -0.38868473(Total Population) +0.58210948(Total White) +0.55858828(Total Black) +0.05514001(Total Hispanic) +0.51284419(Total Other) -0.29137162(Total Number of Households Above Poverty)

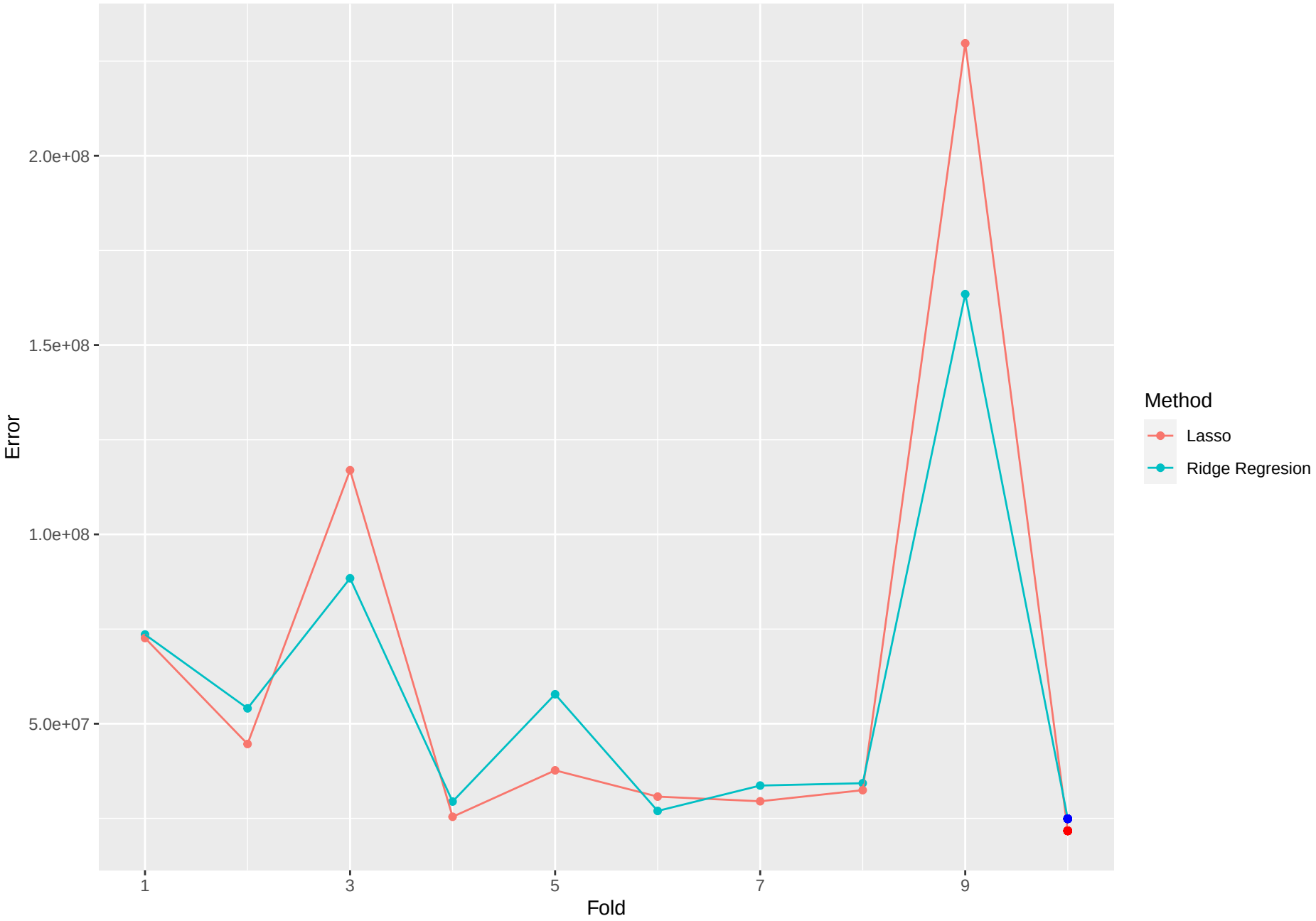
Section 4

Shrinkage Methods

Ridge Regression and Lasso

Recall that Ridge Regression performs better when the response is a function of many predictors of roughly equal size and Lasso performs better when the response is a function of a small subset of predictors.

10-fold Cross Validation was performed to find the optimal tuning parameter λ for each method.


2.0e+08
1.5e+08
1.0e+08
5.0e+07
Error
1
3
5
7
9
Fold
Method
Lasso
Ridge Regresion

Ridge With Lambda that Produces Smallest C.V. Error

The smallest Cross Validation Error for Ridge Regression is 24901393 with and average of 58650401.

The model selected is:

Cumulative Cases = 2659.26735391 -1.46826014(Population Density) +0.01675252(Total Population) -60.01296350(Median Age) +0.03389108(Total White) +0.04046944(Total Black) +0.06843143 (Total Hispanic) +0.03043294 (Total Other) +0.29137162(Total Number of Households Above Poverty) +0.27874317(Total Number of Households Below Poverty)

Lasso With Lambda that Produces Smallest C.V. Error

The smallest Cross Validation Error for Lasso is 21711998 with and average of 64144413.

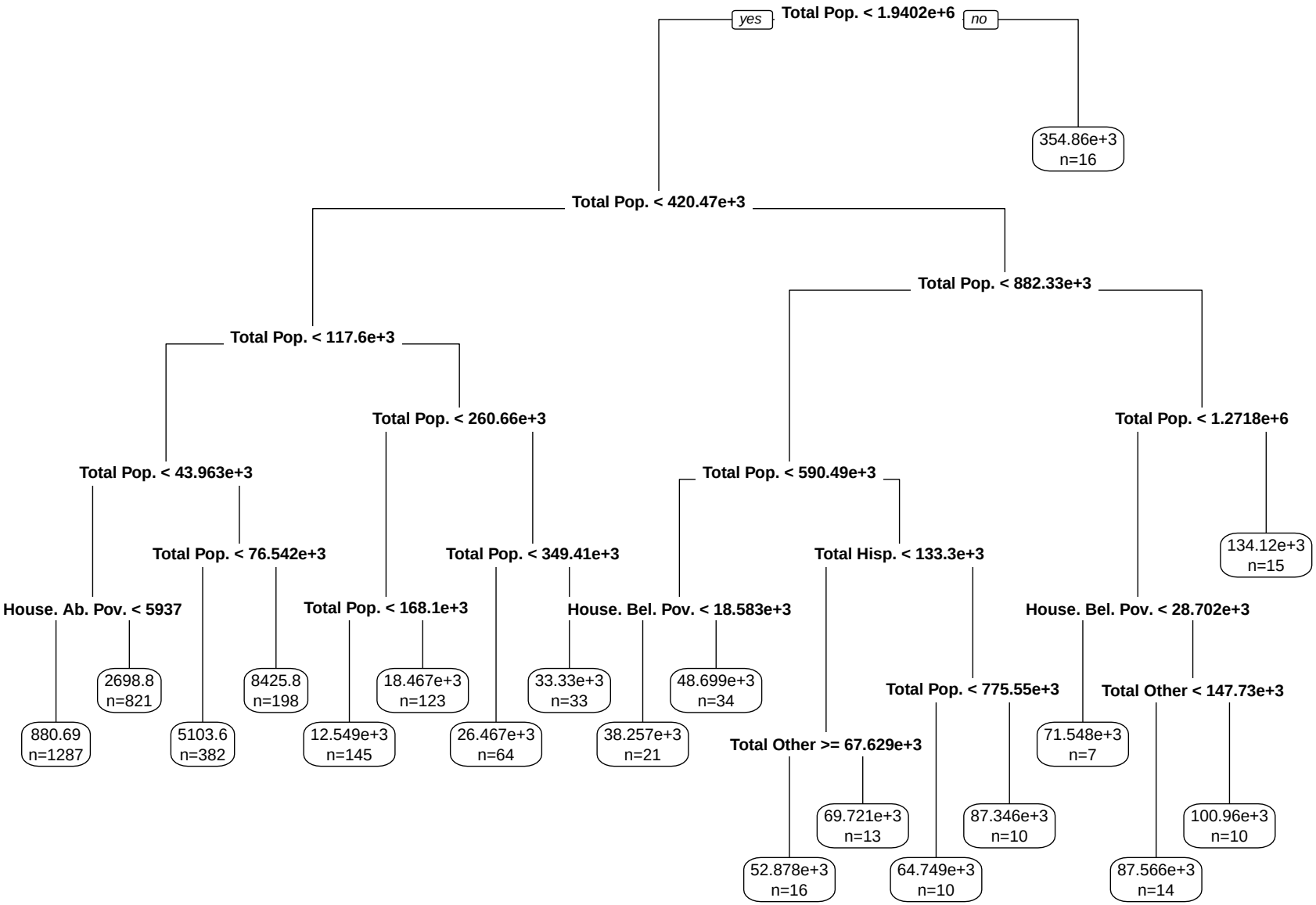
The model selected is:

Cumulative Cases = 286.95400883 +0.04444873(Total Population) +0.02943025(Total White) +0.01685243(Total Black) +0.08547591(Total Hispanic) +0.09854572(Total Number of Households Below Poverty)

Section 5

Tree Based Methods

Regression Trees



Section 6

Classification Analysis

Classification Analysis

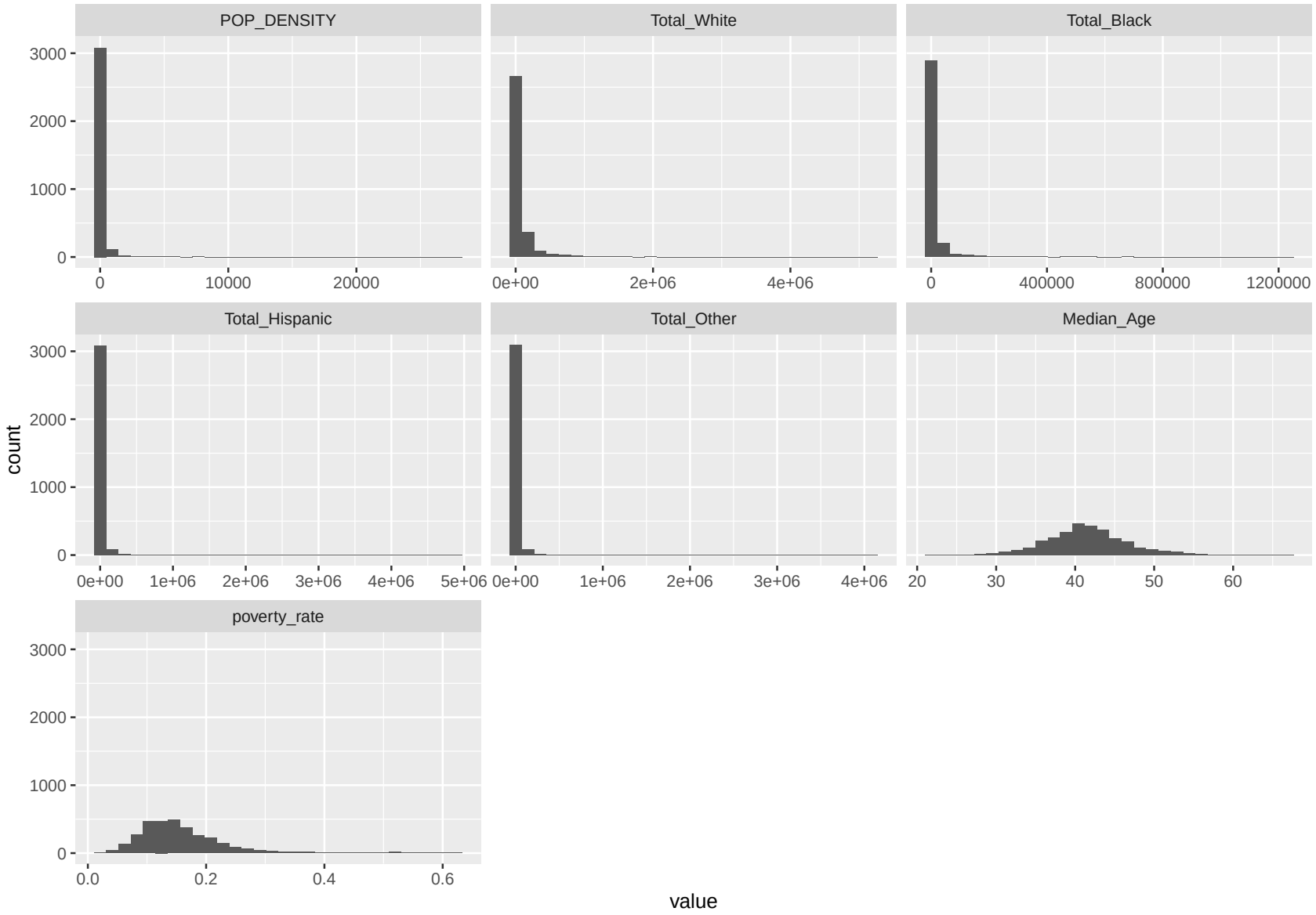
Question: Does a given county have an instance rate higher than the median of the instance rates?

- Binary qualitative response:

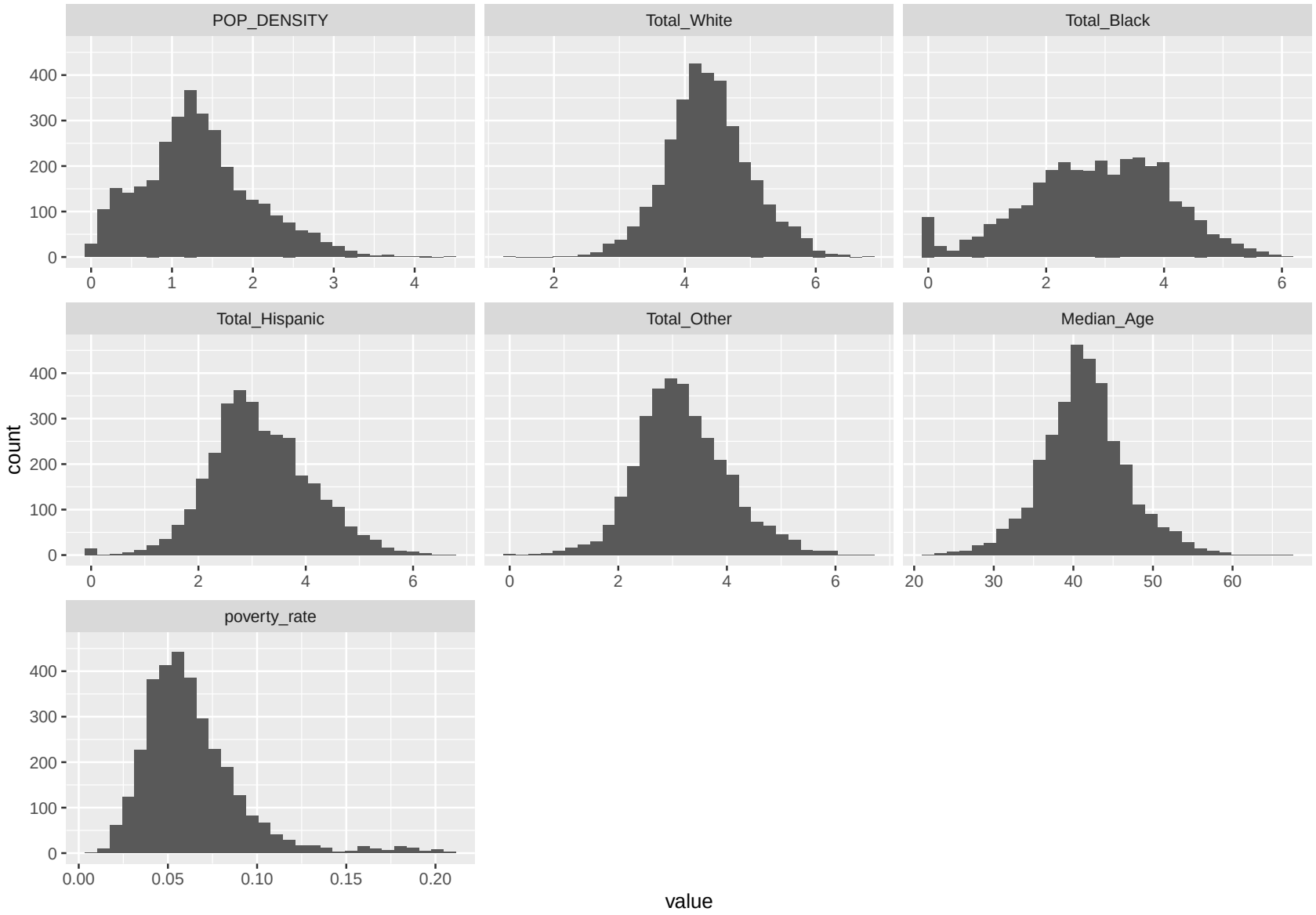
$$\text{Instance rate} = \begin{cases} 1, & \text{above the median of the instance rate} \\ 0, & \text{otherwise} \end{cases}$$

- Consider 'Poverty Rate' instead of 'Total_Households_Above_Poverty' and 'Total_Households_Below_Poverty'

Revisit Plot of Predictors



Data Transformation



Multiple Logistic Regression

- Estimated coefficients of the logistic regression model with all predictors

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.2209	0.6202	10.0300	0.0000
POP_DENSITY	-0.6632	0.1098	-6.0418	0.0000
Total_White	0.3744	0.1544	2.4251	0.0153
Total_Black	0.2020	0.0546	3.7021	0.0002
Total_Hispanic	-0.0332	0.0912	-0.3643	0.7157
Total_Other	-0.4549	0.1185	-3.8391	0.0001
Median_Age	-0.1354	0.0090	-14.9629	0.0000
poverty_rate	-6.1830	1.6370	-3.7772	0.0002

- 'Total_Hispanic' is not statistically significant.

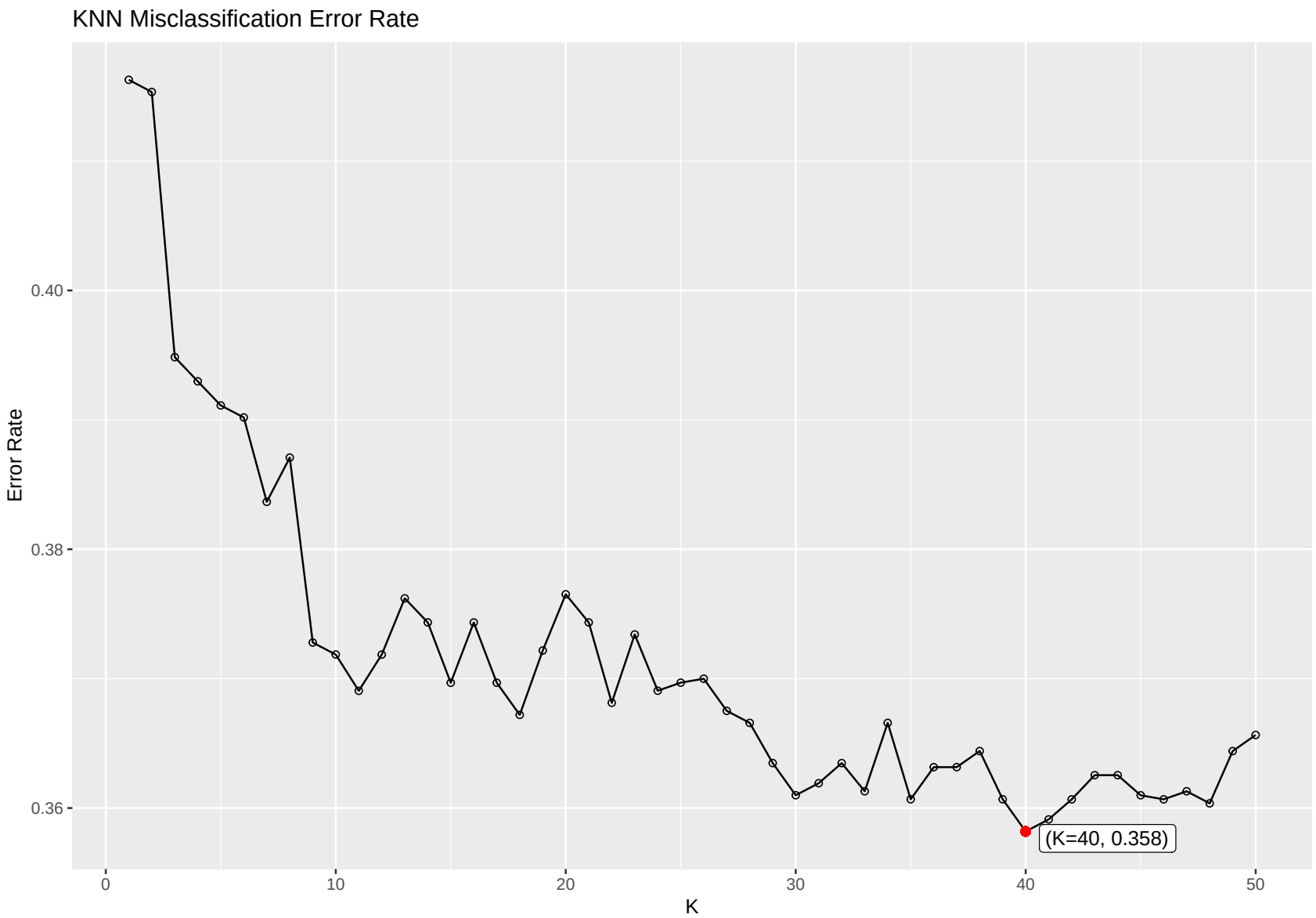
Misclassification Error Rate

- Logistic Regression, LDA, QDA, KNN, and Decision Trees are performed.
- The 5-fold cross-validation is used for each classifier.

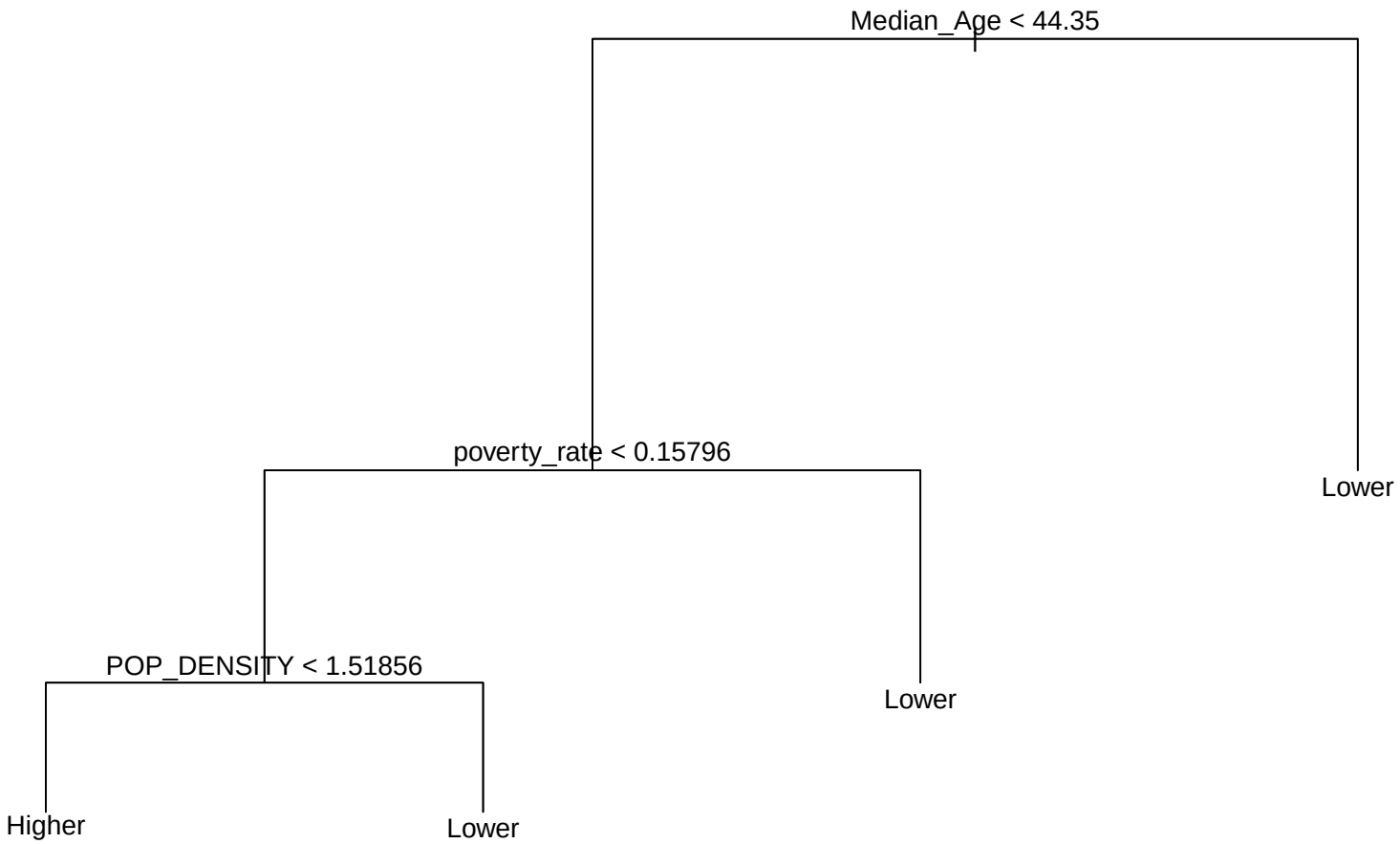
```
glm.misclassification = 0.3694

LDA.misclassification = 0.3675

QDA.misclassification = 0.3719
```

- These all three classifiers have about 63% of overall correct prediction rates.

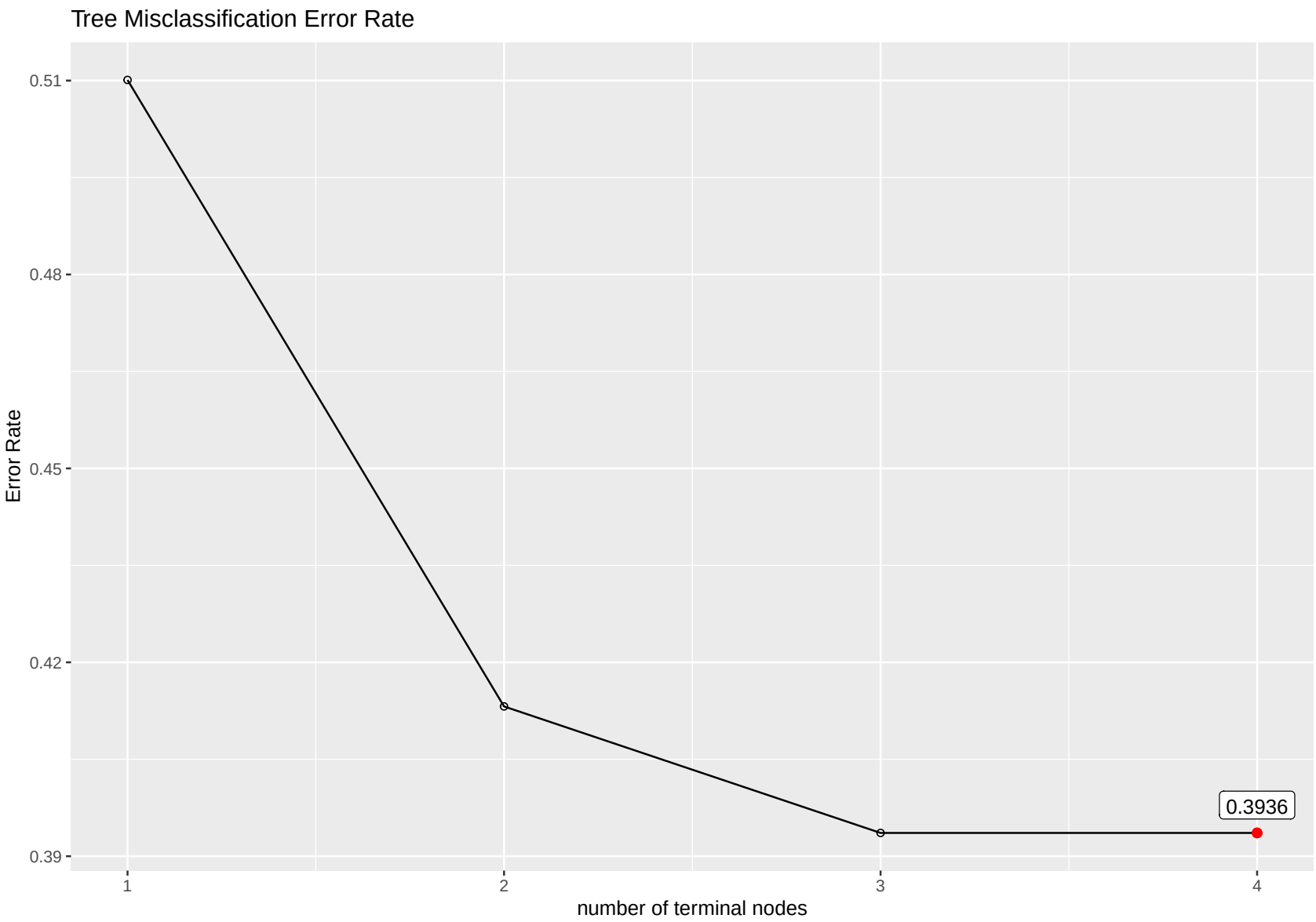
K-Nearest Neighbors (KNN)



Tree-based Classification



Tree Misclassification Error Rate



Comparison of Error Rates

Table 1: Misclassification Error Rate

GLM	LDA	QDA	KNN(K=40)	Tree
0.3694	0.3675	0.3719	0.3582	0.3936

- All of the classifiers produce similar error rates.

Section 7

Conclusion

Results

- Selection methods without cross-validation each selected the same model at each parameter level, however their cross validation had Backward Stepwise selecting a model with more parameters.
- Ridge Regression performed better on average but Lasso had a slightly smaller minimum error during cross validation.
- In prediction of whether a given county is going to have a higher instance rate than the its median, all the classifiers employed in this project (Logistic regression, LDA, QDA, KNN, and Decision Tree) result in the similar accuracy of prediction; about 60% - 63% of overall correct prediction rates.