Investigating Relationships between Demographics and Patient Outcomes in Californian Hospitals

STAT-427/627 Spring 2024 Final Project

Roland Abi Elise Buellesbach Amanda Concepcion
Spencer Grewe

2024-04-25

Table of Contents

1 Executive Summary

In this paper, we dive into California's medical system. Almost 40 million people call California home, as such understanding the efficacy of the California hospital system proves relevant. In this study, we seek to gain a better understanding of the factors that are associated with in-patient deaths in the CA hospital system. Increased staffing, according to the NIH, is associated with decreased levels of in-patient death. Based on this concept, we want to investigate other demographic and economic factors that may be impacting in-patient health.

To start, we completed a series of three regression analyses. In this analysis we want to know if we can predict in-patient deaths due to heart disease per thousand people. We start with a multiple linear regression that provides an unsatisfactory model with only two significant predictor variables. Next, we completed a LASSO regression which yielded insight into how the median income and age of the population impact deaths due to heart disease. The final regression model, a random forest model, reinforces the results from our other models. Through all three models, the number of cases of heart disease and the hospital rating are the most prevalent predictors. Age also continues to be a relevant variable with the population over the age of 55 most at risk. Based on weighting the importance of providing a detailed and robust model with the costs of adding additional variables and the risk of over fitting the data, we recommend using the random forest model.

Next, we turn to a slightly different question regarding the CA hospital system. We complete three classification analyses to try to identify hospitals in Health Professional Shortage Areas (HPSA) based on demographics and geographic location details. To start, we conducted a K-fold model which produced a reduced model based on geographic and population factors. Next, we created a Tree model for the classification where we recommend the usage of the pruned tree. Finally, we used a boost model that produced similar results to the K-fold model. In the end, we recommend the cross-validated boosting model with a shrinkage parameter.

Through this analysis, we gain insight into the California hospital system. We aim to build prediction models for both heart disease in-patient deaths and Health Professional Shortage Areas. HPSA is associated with higher mortality rates of in-patients and thus understanding outcomes in the hospital as well as staffing together provides a broader picture of California hospitals.

2 Data Overview

Data preparation code and documentation was omitted from this report, but can be found within the source code in the GitHub repository under `Data_Clean.qmd`. For this project, we used four data sets.

The "hpsa.csv" file used Health Professional Shortage Areas (HPSA) data extracted from the California Department of Health Access. This data describes the geographic Health

Professional Shortage Area (HPSA) federal designations for Primary Care, Mental Health, and Dental Health in counties within California. This original dataset contains 7,442 observations, and we extracted 10 of the 65 original variables. After identifying repeated counties, we removed repeat observations and selected only those that were unique, which led to a dataset with 587 unique observations. Also, two variables were collapsed: Health Professional Shortage Area (HPSA) population type was collapsed from two categories, migrant population and native Americans; rural status was collapsed from non rural, unknown, partially rural, and rural. This led us to have a total of 8 variables including metropolitan_indicator, designation_type, hpsa_status, hpsa_score, hpsa_designation_population, u_s_mexico_border_county_indicator, rural_status, and hpsa_population_type. We also collected data from the 2020 census and from the at the county level in California and the California State Association of Counties to integrate additional, up-to-date demographic variables into the analysis, such as race, poverty and median income, and proportion of the population within a given age bin.

Our "CA_datapile.xlsx" filed used data from a published dataset in the California's Department of Health Care Access and Information. This dataset, with information from 2016 to 2021, looks at California Hospital Inpatient Mortality rates for 6 medical conditions and 5 procedures. It also contains demographic data for California counties which will let us ask research questions about how county-level demographics intersect patient outcome, specifically mortality. Specifically, it looks at various variables including sex, race, poverty, unemployment, and median income. To conduct regression analysis, a quantitative variable was needed. As a result, our team decided to take the Inpatient Mortality Indicators (IMI) data, extract cases and deaths due to heart attack in a hospital, and standardized these values by calculating the number of heart attack cases for every 1000 people and the number of heart attack deaths for every 1000 people. Strings were transformed to factor variables and we extracted features using regex. Each hospital had a categorical performance rating (worse, better, and as expected). The thousands of observations of hospital, cases, and demographic data was distilled to about a hundred entries. Every row is a unique combination of the cases and death rates per 1000 in hospitals in a specific county with a particular rating, as well as the additional hospital HPSA status and demographic information. Also, we removed hospital observational rows with missing entries in hospital quality, which is a factor that can lead to bias. To address missing data, we could have explored various alternative tactics: mean imputation and multivariate imputation, and future studies may address these concerns.

3 Classification Models

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
```

```
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

3.1 Method 1: Logistic Regression

3.1.1 Training and Testing

Here we will be using 60% of our dataset for training and 40% for testing

3.1.2 Full model

```
# A tibble: 13 x 5
   term                                    estimate std.error statistic p.value
   <chr>                                      <dbl>     <dbl>     <dbl>   <dbl>
 1 (Intercept)                              2.55e+1   7.30e+3   3.50e-3 0.997
 2 metropolitan_indicatorMetropolitan      -1.02e+0   7.49e+3  -1.36e-4 1.00
 3 metropolitan_indicatorNon-Metropolitan  -7.54e-1   7.55e+3  -9.98e-5 1.00
 4 metropolitan_indicatorUnknown           -2.09e+1   7.30e+3  -2.87e-3 0.998
 5 designation_typeHigh Needs Geographic H~ -3.71e-1  5.27e-1  -7.05e-1 0.481
 6 designation_typeHPSA Population         -2.18e+0   1.10e+0  -1.98e+0 0.0474
 7 hpsa_score                              -1.81e-1   5.90e-2  -3.08e+0 0.00209
 8 hpsa_designation_population              1.93e-6   6.64e-6   2.91e-1 0.771
 9 u_s_mexico_border_county_indicatorY     -8.40e-2   7.90e-1  -1.06e-1 0.915
10 rural_statuspartially_rural              3.48e-2   5.73e+3   6.07e-6 1.00
11 rural_statusrural                        7.71e-2   4.16e-1   1.85e-1 0.853
12 rural_statusunknown                     -1.81e+0   8.54e-1  -2.12e+0 0.0343
13 hpsa_population_typenative_american     -3.20e+0   1.07e+0  -2.99e+0 0.00277
```

3.1.3 Reduced model

```

Call:
glm(formula = as.factor(hpsa_status) ~ designation_type + hpsa_score +
    rural_status + hpsa_population_type, family = "binomial",
    data = hpsa_train)

Coefficients:
                                           Estimate Std. Error z value
```

```
(Intercept)                                  6.25862    0.93204    6.715
designation_typeHigh Needs Geographic HPSA  -0.64684    0.45256   -1.429
designation_typeHPSA Population             -1.46585    0.60276   -2.432
hpsa_score                                  -0.31939    0.03875   -8.242
rural_statuspartially_rural                 13.09281  759.83452    0.017
rural_statusrural                           -0.05037    0.33005   -0.153
rural_statusunknown                         -2.66319    0.83559   -3.187
hpsa_population_typenative_american         -2.41549    0.59263   -4.076
                                           Pr(>|z|)
(Intercept)                                1.88e-11 ***
designation_typeHigh Needs Geographic HPSA  0.15293
designation_typeHPSA Population             0.01502 *
hpsa_score                                  < 2e-16 ***
rural_statuspartially_rural                 0.98625
rural_statusrural                           0.87871
rural_statusunknown                         0.00144 **
hpsa_population_typenative_american        4.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 453.02  on 351  degrees of freedom
Residual deviance: 249.95  on 344  degrees of freedom
AIC: 265.95

Number of Fisher Scoring iterations: 15
```

3.1.4 Compare models Full and reduced logistic regression models

Accuracy and Prediction Error Estimates Logistic regression models

models	Accuracy	Error	Adjusted_MSEP
Reduced	0.812766	0.187234	0.1995639
Full	0.893617	0.106383	0.1497533

3.1.5 Summary of Logistic regression models

The reduced model has an accuracy of 0.812766 and a mean prediction error rate of 0.1995639 while the full model has an accuracy of 0.893617 which represent an 8% increase in accuracy

Table 1: Model comparison

```
Analysis of Deviance Table

Model 1: as.factor(hpsa_status) ~ designation_type + hpsa_score + rural_status +
    hpsa_population_type
Model 2: as.factor(hpsa_status) ~ metropolitan_indicator + designation_type +
    hpsa_score + hpsa_designation_population + u_s_mexico_border_county_indicator +
    rural_status + hpsa_population_type
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       344     249.95
2       339     191.42  5   58.529 2.446e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

over the reduced model and a mean prediction error rate of 0.1497533. The full model also has a lower adjusted mean square error of prediction 0.1497533 compared to the reduced model 0.1995639 using 10-fold cross validation **?@tbl-accuracy-preds** .

Giving the results obtained and from from the reduced and full logistic regression models and the results of model comparison we lack sufficient evidence (p = 2.446e-11) Table 1 to conclude that logistic regression (reduced) model is better than the logistic regression (full) model. Consequently, we will recommend the full logistic regression model as the better model with predictors such as designation:HPSA_population, hpsa score rural_status:unknown and hpsa_population_type:native american among the significant predictors in determining hpsa status.

3.2 Method 2: Decision Tree

3.2.1 Un-pruned decision Tree

3.2.2 Pruned Tree

```
$size
[1] 12 11 10  8  3  2  1

$dev
[1]  69  69  63  61  62  77 100

$k
[1] -Inf  0.0  1.0  1.5  3.0 13.0 25.0
```

Table 2: Unpruned model output

```
Classification tree:
tree(formula = as.factor(hpsa_status) ~ ., data = train)
Variables actually used in tree construction:
[1] "hpsa_score"                  "hpsa_designation_population"
Number of terminal nodes:  12
Residual mean deviance:  0.6527 = 161.9 / 248
Misclassification error rate: 0.1385 = 36 / 260
```

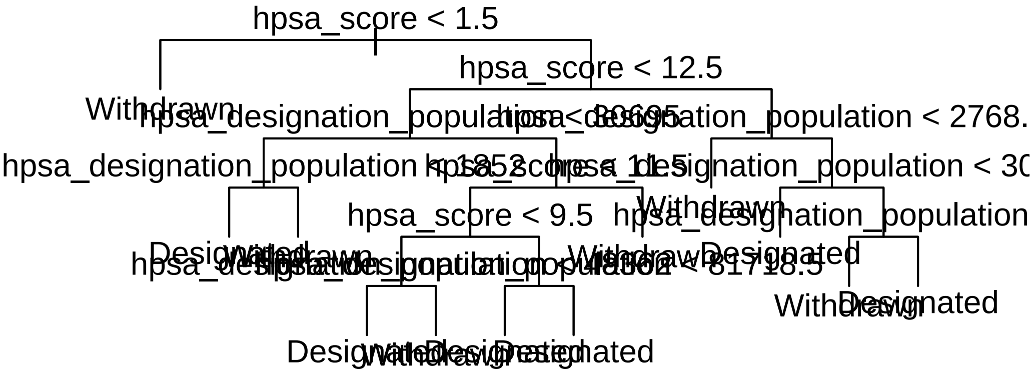

hpsa_score < 1.5
hpsa_score < 12.5
Withdrawn
hpsa_designation_population < 30695
hpsa_designation_population < 2768.
hpsa_designation_population < 1852
hpsa_score < 11.5
hpsa_designation_population < 30
hpsa_score < 9.5
Withdrawn
hpsa_designation_population
Designated
Withdrawn
hpsa_designation_population
hpsa_designation_population
Withdrawn
hpsa_designation_population < 81718.5
Designated
Designated
Withdrawn
Designated
Designated
Withdrawn
Designated

Figure 1: Decision tree from unpruned model with 12 nodes

```
$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

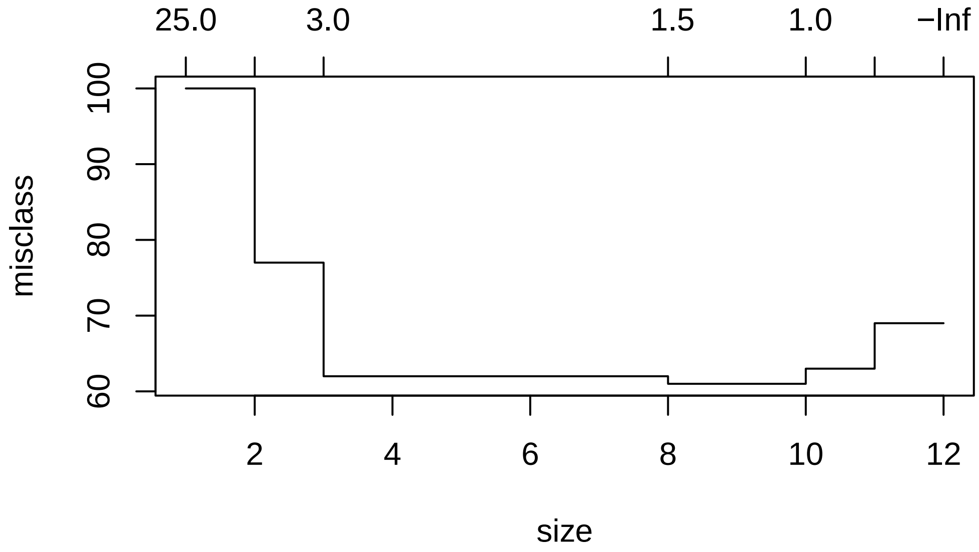


Figure 2: Misclassification rate from Cross valid

Tree size 8 corresponds to the lowest cross-validated classification error rate with the fewest nodes.

3.3 Comparing the performance of the tree model

Accuracy and Prediction Error Estimates Tree models

models	Accuracy	Error
Unpruned	0.7951070	0.2048930
Pruned	0.7981651	0.2018349

Table 3: Pruned model output

```
Classification tree:
snip.tree(tree = tr, nodes = c(53L, 12L, 15L))
Variables actually used in tree construction:
[1] "hpsa_score"                  "hpsa_designation_population"
Number of terminal nodes:  8
Residual mean deviance:  0.7143 = 180 / 252
Misclassification error rate: 0.1538 = 40 / 260
```

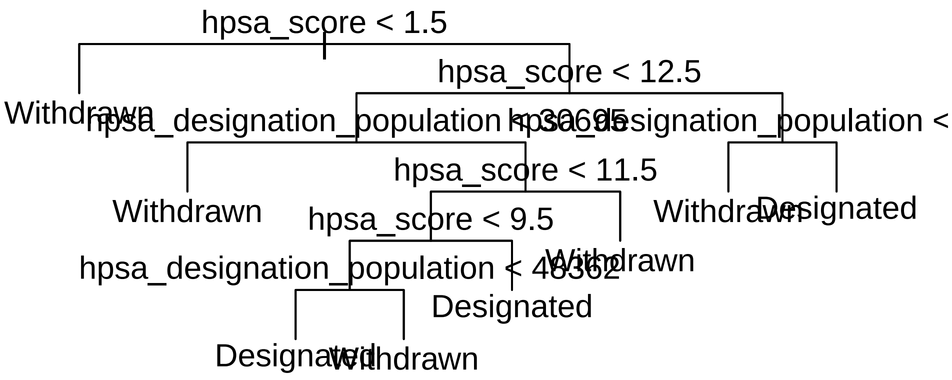


Figure 3: Decision tree from pruned model with 8 nodes

3.3.1 Summary of Decision Tree models

The result obtained shows that the unpruned model with a training error rate of 0.1385 had twelve terminal nodes and retained the variables hpsa score and hpsa designation population in the model. Meanwhile the pruned model using cross validation to obtain the number of trees with the minimum misclassification rate Figure 2 had a slightly higher training error rate as shown in Table 3 cross validated model with eight terminal nodes.

"HPSA Score" and HPSA designation population were the dominant predictors in both the unpruned and pruned models. Overall, The pruned model had a mean prediction error rate of 0.2018349 which is a little lower than the unpruned tree which has an error rate of 0.204893. Meanwhile the accuracy of the pruned tree is 0.7981651 which represent a 0.3% increase in accuracy over the unpruned tree model with accuracy 0.795107 . We recommend the pruned tree Figure 3 model.

3.4 Method 3: Boosting Method

3.4.1 Initial boosting model

```
gbm(formula = hpsa_status ~ ., distribution = "Bernoulli", data = train,
    n.trees = 3000)
A gradient boosted model with Bernoulli loss function.
 3000 iterations were performed.
There were 7 predictors of which 7 had non-zero influence.
```

```
                                                                  var
metropolitan_indicator                         metropolitan_indicator
hpsa_score                                                 hpsa_score
hpsa_population_type                             hpsa_population_type
hpsa_designation_population               hpsa_designation_population
designation_type                                     designation_type
rural_status                                             rural_status
u_s_mexico_border_county_indicator u_s_mexico_border_county_indicator
                                        rel_inf
metropolitan_indicator             76.228639138
hpsa_score                         18.837306324
hpsa_population_type                2.689845808
hpsa_designation_population         1.149605087
designation_type                    0.944460900
rural_status                        0.142845378
u_s_mexico_border_county_indicator  0.007297366
```

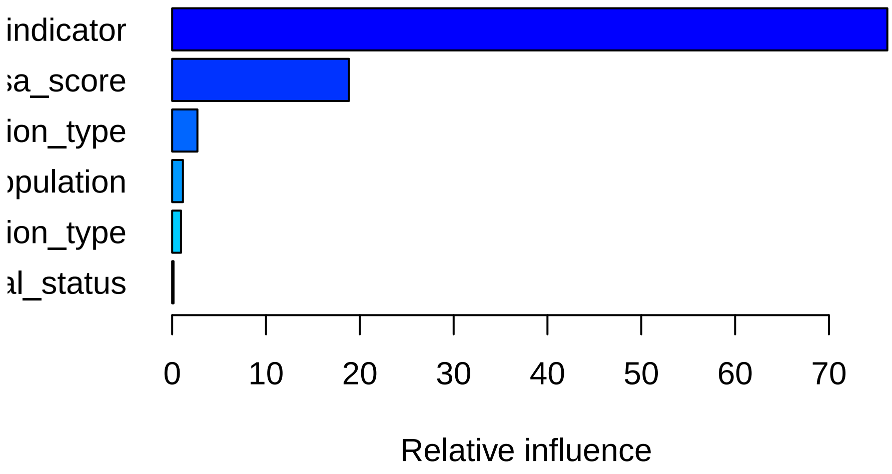


Figure 4: Top 6 important variables for initial boosting model

3.4.2 Using Shrinkage parameter

```
gbm(formula = hpsa_status ~ ., distribution = "Bernoulli", data = train,
    n.trees = 5000, shrinkage = 0.001)
A gradient boosted model with Bernoulli loss function.
 5000 iterations were performed.
There were 7 predictors of which 7 had non-zero influence.


                                                                  var
metropolitan_indicator                         metropolitan_indicator
hpsa_score                                                 hpsa_score
hpsa_population_type                             hpsa_population_type
hpsa_designation_population               hpsa_designation_population
designation_type                                     designation_type
rural_status                                             rural_status
u_s_mexico_border_county_indicator u_s_mexico_border_county_indicator
                                       rel_inf
metropolitan_indicator             71.37882009
hpsa_score                         19.27404814
hpsa_population_type                4.00518646
hpsa_designation_population         3.41247481
```

```
designation_type                  1.19690539
rural_status                      0.70524554
u_s_mexico_border_county_indicator  0.02731957
```

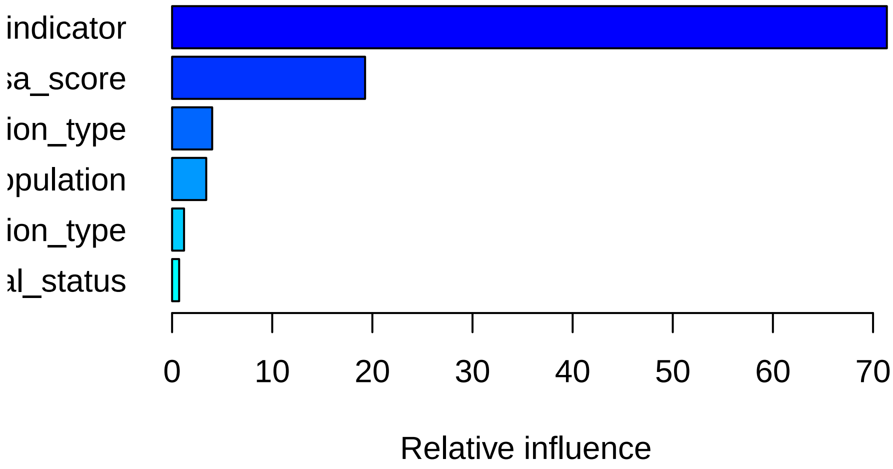


Figure 5: Top 6 important variables for boosting with shrinkage model

3.4.3 Using Cross Validation with shrinkage parameter

```
gbm(formula = hpsa_status ~ ., distribution = "Bernoulli", data = train,
    n.trees = 3000, shrinkage = 0.001, cv.folds = 10)
A gradient boosted model with Bernoulli loss function.
 3000 iterations were performed.
The best cross-validation iteration was 3000.
There were 7 predictors of which 7 had non-zero influence.

Cross-validation confusion matrix:
   0   1 Pred. Acc.
0 85  15       85.0
1 33 160       82.9

Cross-validation prediction Accuracy = 83.62%
```

```
                                                                    var
metropolitan_indicator                           metropolitan_indicator
hpsa_score                                                   hpsa_score
hpsa_population_type                               hpsa_population_type
hpsa_designation_population                 hpsa_designation_population
designation_type                                       designation_type
rural_status                                               rural_status
u_s_mexico_border_county_indicator u_s_mexico_border_county_indicator
                                        rel_inf
metropolitan_indicator             76.228639138
hpsa_score                         18.837306324
hpsa_population_type                2.689845808
hpsa_designation_population         1.149605087
designation_type                    0.944460900
rural_status                        0.142845378
u_s_mexico_border_county_indicator  0.007297366
```

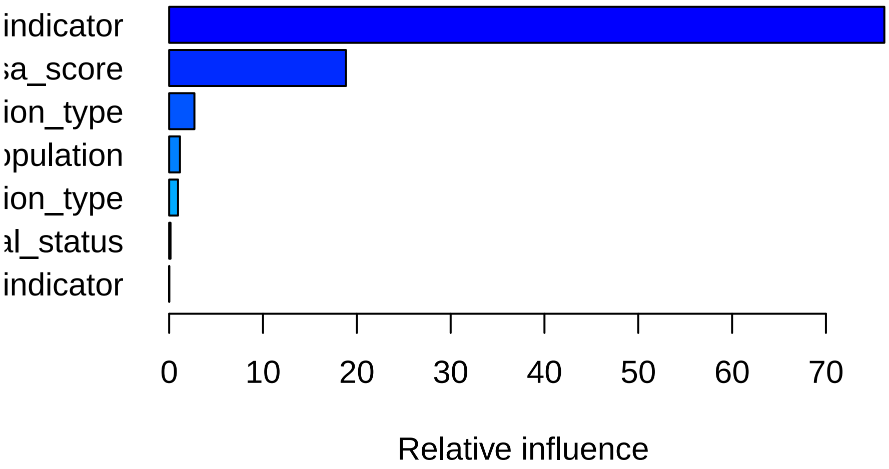
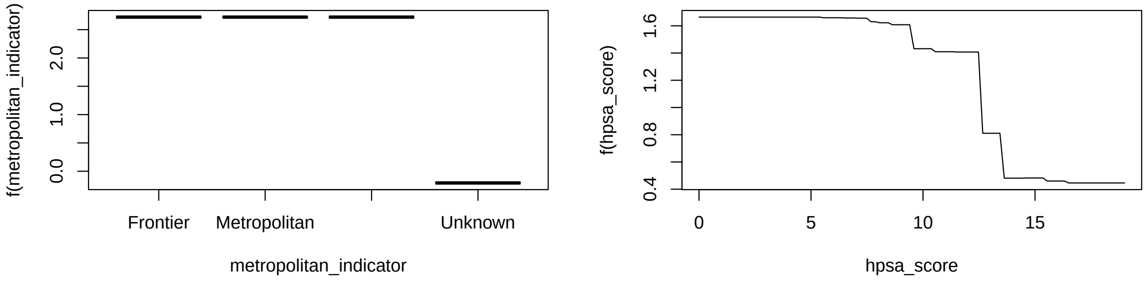
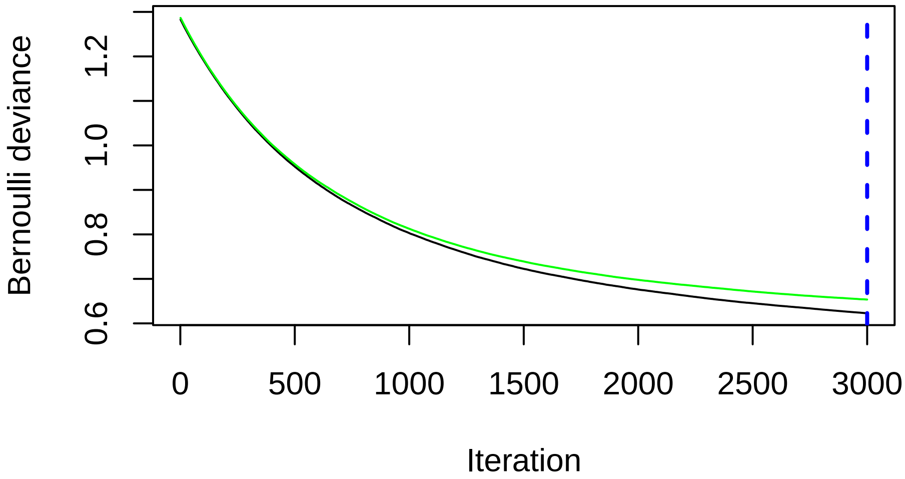


Figure 6: Top 6 important variables for CV boosting_shrinkage model



(a) Marginal effect of top two predictors

(b) Marginal effect of top two predictors



```
[1] 3000
```

Both training deviance (black line) and testing deviance (green line) seemed fairly closed.

Accuracy and Prediction Error Estimates of Boosting models

Models	Accuracy	Error
initial model	0.877551	0.122449
boost_shrinkage	0.877551	0.122449

boost_shrinkage_cv	0.877551	0.122449

3.4.4 Summary of boosting models

Our analysis shows that increasing or decreasing the shrinkage rate did not affect the performance of the model with shrinkage parameter, and the cross validated model with shrinkage parameter .The cross-validated boosted model with shrinkage parameter **?@tbl-boostingshrinkage-cv** , boosted model with shrinkage parameter **?@tbl-boostingshrinkage** and initial boosted **?@tbl-initialboosting** model all have similar prediction mean squared error. Top important predictors in the initial Figure 4 and boost_shrinkage Figure 5 models were metropolitan indicator and hpsa score. While the top important predictors in the boost_shrinkage_cv Figure 6 model were hpsa designation population, hpsa status and metropolitan indicator. In all three models, hpsa score and metropolitan indicator were the consistent predictors of hpsa status. We will recommend the boost_shrinkage_cv model because we believe it will likely generalize to unseen data having used cross validation to tune model parameters.

4 Regression Models

4.1 Method 1. Multiple Linear and Ridge Regression

Here we are working on a basic multiple linear model with cross validation.

Question: Can mortality rate from heart disease per 1000 people be predicted from demographic conditions within counties in California?

To start, we will employ a series of multiple regressions to get a better understanding of the relationships at play and which variables are most useful in predicting the mortality rate from heart disease.

```
Rows: 99 Columns: 40
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): County, Rating
dbl (38): TotalPop, MalePop, FemalePop, MedianAge, MaleMedianAge, FemaleMedi...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

When I initially began building this model, I ran into issues in calculating the MSE because of a couple of NA values. To prevent this issue, I checked the data for NA values. Both PercFemale and PercUnder18yo have NA values and were thus removed from the full regression model.

4.1.1 Full Model

Looking at the full model, we see that the full model is statistically significant with a p-value of almost zero. Further, it has an adjusted R^2 value of 0.9345 telling us that the model explains 93.45% of the variation in the results. This is a strong result. But, only a handful of the variables are significant in the model. ::: {.cell} ::: {.cell-output .cell-output-stdout}

```
Call:
lm(formula = DeathsperThou ~ TotalPop + MalePop + FemalePop + 
    MedianAge + MaleMedianAge + FemaleMedianAge + Under5yoPop + 
    Under18yoPop + Pop21andOlder + Pop55andOlder + Pop60andOlder + 
    Pop65andOlder + num_desig + amer_indian + black + hispanic + 
    asian + multi_race + pac_island + white + poverty + labor_force + 
    unemployed + median_income + TotalDeaths + TotalCases + prop_hpsa + 
    nonwhite_prop + Rating + Deaths + unempl_prop + poverty_prop + 
    prop55older + Cases + HospNumber + CasesperThou + female_prop, 
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26702 -0.02710  0.00018  0.02693  0.38361 

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)   
(Intercept)      1.476e+01  9.677e+00   1.525  0.13482   
TotalPop        -5.931e-05  3.533e-05  -1.679  0.10079   
MalePop         -3.238e-05  2.615e-05  -1.238  0.22266   
FemalePop               NA         NA      NA       NA   
MedianAge        3.873e-01  2.140e-01   1.810  0.07769 . 
MaleMedianAge   -3.582e-01  1.234e-01  -2.904  0.00592 **
FemaleMedianAge -4.030e-01  1.508e-01  -2.672  0.01076 * 
Under5yoPop     -4.857e-05  7.834e-05  -0.620  0.53872   
Under18yoPop     8.693e-05  5.331e-05   1.631  0.11062   
Pop21andOlder    9.340e-05  5.649e-05   1.653  0.10588   
Pop55andOlder   -6.930e-05  6.245e-05  -1.110  0.27362   
Pop60andOlder    1.896e-05  1.107e-04   0.171  0.86482   
Pop65andOlder    4.339e-05  7.970e-05   0.544  0.58910   
```

```
num_desig       -2.381e-03  2.080e-03  -1.145  0.25892    
amer_indian      7.318e-06  3.708e-05   0.197  0.84453    
black           -6.558e-06  9.342e-06  -0.702  0.48668    
hispanic         8.807e-06  5.090e-06   1.730  0.09115 .  
asian            4.462e-06  4.295e-06   1.039  0.30494    
multi_race       6.853e-05  5.490e-05   1.248  0.21902    
pac_island      -2.130e-05  6.064e-05  -0.351  0.72719    
white            1.295e-06  4.764e-06   0.272  0.78711    
poverty          5.334e-06  5.782e-06   0.922  0.36167    
labor_force     -1.302e-05  8.837e-06  -1.474  0.14815    
unemployed      -1.713e-05  1.891e-05  -0.906  0.37047    
median_income    2.027e-05  1.331e-05   1.523  0.13534    
TotalDeaths     -2.665e-04  9.168e-04  -0.291  0.77275    
TotalCases      -5.199e-05  1.883e-05  -2.761  0.00858 ** 
prop_hpsa       -2.052e-02  2.950e-01  -0.070  0.94489    
nonwhite_prop   -1.694e+00  1.696e+00  -0.999  0.32367    
RatingBetter    -1.170e-01  1.015e-01  -1.153  0.25571    
RatingWorse     -4.386e-02  1.007e-01  -0.436  0.66537    
Deaths           2.238e-04  1.519e-04   1.473  0.14839    
unempl_prop      4.769e+00  4.009e+00   1.190  0.24105    
poverty_prop    -4.411e+00  6.737e+00  -0.655  0.51629    
prop55older      2.768e+01  1.998e+01   1.386  0.17337    
Cases           -3.037e-06  3.359e-06  -0.904  0.37129    
HospNumber      -2.326e-04  1.308e-04  -1.778  0.08275 .  
CasesperThou     3.264e-02  3.848e-03   8.482 1.46e-10 ***
female_prop     -1.737e+01  1.094e+01  -1.587  0.12011    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.102 on 41 degrees of freedom
Multiple R-squared:  0.9655,    Adjusted R-squared:  0.9345 
F-statistic: 31.06 on 37 and 41 DF,  p-value: < 2.2e-16
```

::: :::

When looking at plots for this model, some concerns arise. For example, the plotting on the QQ-plot does not demonstrate a strong linear relationship. The edges of the plot deviate from the line suggesting that the model is not perfectly linear. Further, the scattered dots on the Leverage plot show several points on the line of Cook's Distance signifying that they have overstated influence on the model. Further attentions is paid to these plots in the appendix.

```
[1] 0.07419842
```

The MSE is .0742. This is a low MSE in general. Next, we can see how a reduced model will work and then we will have a good point of comparison between the two models.

Here, I will take only the significant variables from the above model for the reduced model. This includes variables that are significant up to the p=.1 threshold This was chosen because so few variables are significant and I want to see what this model will yield. Then, we can remove more variables if the results warrant that action.

4.1.2 Reduced Model

```
Call:
lm(formula = DeathsperThou ~ MedianAge + MaleMedianAge + FemaleMedianAge + 
    hispanic + CasesperThou + TotalCases + HospNumber, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40557 -0.05661 -0.01361  0.05143  0.56690 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.900e-01  1.517e-01  -1.252   0.2146    
MedianAge        2.057e-01  1.394e-01   1.476   0.1443    
MaleMedianAge   -1.208e-01  7.612e-02  -1.588   0.1168    
FemaleMedianAge -8.090e-02  6.746e-02  -1.199   0.2344    
hispanic         3.747e-08  8.951e-08   0.419   0.6768    
CasesperThou     3.500e-02  1.482e-03  23.615   <2e-16 ***
TotalCases      -6.504e-07  1.326e-06  -0.491   0.6253    
HospNumber      -1.122e-04  4.518e-05  -2.483   0.0154 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1192 on 71 degrees of freedom
Multiple R-squared:  0.9185,	Adjusted R-squared:  0.9104 
F-statistic: 114.3 on 7 and 71 DF,  p-value: < 2.2e-16
```

This model yields a p-value of almost zero. It has an adjusted R^2 value of 0.9104 which is lower than the previous value from the full model of 0.9345. However, the decrease in the number of variables might be worth the slightly lower R^2 value.

```
[1] 0.05392615
```

However, we see a lower MSE at 0.0539 compared to 0.0742 from the full model. However, `hispanic`, `MedianAge`, `MaleMedianAge`, `FemaleMedianAge`, and `TotalCases` are no longer significant in this model.

To get an even better understanding of the data, we reduce the model again to just the significant variables in the reduced model. We are down to just two predictor variables, `CasesperThou` and `HospNumber`. This makes logical sense that the number of cases per thousand and the ranking of the hospital would be good predictor variables of the deaths per thousand.

```
Call:
lm(formula = DeathsperThou ~ CasesperThou + HospNumber, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42998 -0.05724 -0.00882  0.02736  0.69795 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.317e-03  1.768e-02   0.527  0.59983    
CasesperThou  3.503e-02  1.451e-03  24.140  < 2e-16 ***
HospNumber   -1.230e-04  4.346e-05  -2.830  0.00595 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.122 on 76 degrees of freedom
Multiple R-squared:  0.9087,	Adjusted R-squared:  0.9063 
F-statistic: 378.1 on 2 and 76 DF,  p-value: < 2.2e-16

[1] 0.05392615
```

This small model yields the same MSE as the previous reduced model of 0.0539 and only requires two variables. We still have a p-value of almost zero. However, we continue to lose some power in the adjusted R-squared with a value of 0.9063 compared to 0.9104 in the reduced model and 0.9345 in the full model.

4.1.3 Comparison Between the Two Models and Recommendation

Between the original two models, the reduced model should be chosen. It offers a more accurate prediction of the rate of deaths due to heart disease per thousand people. The reduced model uses only 7 prediction variables compared to the more than 30 variables that the full model requires. Further, the smallest model with only 2 variables gives the same low MSE value with even fewer predictor variables. This implies that many of those variables are either not useful in prediction or they might have high levels of multi-co linearity This concept will be explored further in the next regression models.

4.2 Method 2. LASSO Regression

4.2.1 Further Exploration of Multicollinearity

In social science and health data, multicollinearity has complex causes and is generally unavoidable. There are a few notable instances of collinearity within the regression predictors (SI Fig. 8). The non-white racial population proportion and the population proportion aged 55 years and older can be reliably predicted from the rest of the data, and certain pairs, such as poverty and unemployment, show evidence of an intuitively strong linear association.

While the cause of these co-correlations is outside of the scope of our analysis, we will now explore two additional regression methods, LASSO and random forest, which use shrinkage or bagging to reduce the variance in the prediction. Unlike simply removing predictors from the analysis, these two methods also allow for unbiased searches across all predictors to construct the best final model.

4.2.2 LASSO Regression

We employed a 10-fold LASSO regression to determine whether deaths per 1000 individuals due to heart disease can be predicted given demographic information, as well as information about the hospitals located within counties in California, controlling for heart disease cases and with a shrinkage penalty applied to address multicollinearity.

Table 4: Overview and prediction performance of the LASSO models at two different L1 penalty strengths.

Model	Prediction MSE	Number of Predictors
lambda 1se	0.0479827	1
lambda min	0.0150106	3

The prediction MSE for the lambda 1se model, which had only one significant predictor, is 0.04798, while the prediction MSE for the lambda min model with three predictors is 0.01501 (Table 4). The sole predictor in the first model was the heart disease case rate, and the difference in prediction error between the two models is a fairly large 68.72%. It is clear that these variables are lending predictive power to the model, but not enough to significantly outweigh a stronger L1 penalty imposed by LASSO.

In the lambda min model, we find that additional predictors of deaths due to heart disease include the median income, hospital rating, number of hospitals in the county at that rating level, and proportion of residents 55 years old and older (SI Table 6). This agrees with common sense: we can imagine that the overall rate of deaths due to heart disease in a given area is due to demographic effects on the health of the general population and economic effects impeding treatment. We also find that the rating of a hospital on its own has no statistically significant effect on heart attack deaths until it interacts with the number of heart attack cases (SI Table 6). In general, subgroups of hospitals rated “Worse” tended to have higher deaths, while deaths were expected to *decrease overall* for hospitals with “Better” ratings with cases remaining the same. LASSO demonstrates that there is a trade-off between model simplicity and prediction power, and only in complex models do we see an effect due to demographic and hospital-level variables.

4.3 Method 3. Regression Trees

4.3.1 Random Forest

Table 5: Summary of cross-validated random forest results.

Metric	Value
Number of Predictors	7
Number of Trees	398
Prediction MSE	0.0299

An exhaustive search using cross validation produced the optimal number of predictors and number of trees to use for the training data: 7 predictors and 398 trees respectively. The random forest model has a prediction MSE of 0.02992, which corresponds to a 37.64% decrease in prediction MSE with respect to the lambda 1se LASSO regression model, and a 44.49% decrease with respect to the best MLR model (Table 5). We find that 81.33% of the variation in the heart disease death rate is explained by the combination of 7 variables, and it is again clear that variables beyond case rate are lending predictive power to the model.

We find that hospital rating and amount are significantly used predictors in the random forest, and likely reflect the quality and accessibility of care in different counties (SI Fig. 9). The effect of hospital rating and the number of hospitals far outweigh the effect of demographic factors, including the proportion of residents aged 55 years and older. In general, the random forest model performs with the second lowest prediction error of all regression models explored, and strikes a reasonable balance between bias and complexity in the number of predictors. We fail to find strong evidence of demographic factors, but these regression models illustrate that that hospital quality and number in counties in the state of California are associated with deaths due to heart disease when controlling for variation in the number of heart disease cases.

5 Ethical Analysis

Managing health-related data and models comes with a wide variety of serious ethical concerns. Understanding the risks of biased data or an inaccurate model is critical to safeguarding against these. As such, as we work on these models, we must keep in mind the privacy of patients, the quality of our data, and its ability to accurately represent the demographics of the whole population, and we must be mindful of the variables that are included or excluded.

1) Managing data: Collecting medical data is a delicate challenge. First and foremost, health data is personal and sensitive. It must be handled with the utmost respect and caution as it contains sensitive and personal information. The US legal system, through The Health Insurance Portability and Accountability Act of 1996 (HIPAA), protects sensitive patient information. However, to study the health system and understand the weaknesses and strengths of the current hospital system, we need data. As such, as data scientists, we must be sure to handle data in a careful manner that ensures that individual and identifiable information is not connected with the data used to build the model. Protecting privacy creates challenges in collecting the data and ensuring that the data is representative. (NIH)

2) Biases and Representation: Medical data must be representative of the population it is measuring or we risk creating an inaccurate and harmful model. Inaccurate data can reinforce stereotypes leading to misdiagnosis and mistreatment. Medical data differs from other data sources because inaccuracies are often a matter of life and death. In this particular set of models, we need to be cautious about the demographics and our ability to generalize the results. For example, this data gives good insight into California's hospital system, however, California is not representative of the nation. California's median age (37.9) is the eleventh youngest in the nation and almost a year younger than the median age nationally (38.8). Further, California is more diverse than other states and has a larger population of immigrants (27%) compared to the rest of the country (12%) (PPIC). Further, the median income in California is more than $15,000 greater than the national median income (US Data). These are a couple of examples where California's data does not represent the greater US. It is important to note these disparities and study how they might impact the data and prevent us from generalizing our data to other hospital systems. (NIH)

3) Inclusion of Variables: In each model, we had to select which variables to include and which to exclude. For example, in the multiple linear regression models, `hispanic` was the only race variable that was significant in the full model and thus was included in the reduced model while others were not. This choice, even based on the data, is one

> that can impact the outcome. Transparency in which variables are included in the final models is critical to the reproducibility of the model and the preventing harmful biases.

Conversely, good data and strong models can serve as guidance to hospitals and be accountability measures. For example, the NIH has found a 10% increase in in-hospital mortality rates due to unsafe practices while increased staffing is correlated with a 12% decrease in deaths. Accurate data reporting can provide accountability for hospital practices. This gives community members insight into the medical system and the needed improvements in their locality.

Further, better public policy can stem from accurate reporting. For example, when considering our models, overlaying deaths and demographic information can highlight at-risk groups that need further support. We find in our LASSO analysis that patients over the age of 55 are at a higher risk of death. As such, policymakers know that they need to further focus on resources for this community. Continued monitoring of the death rate can demonstrate if policy changes yield a change in mortality rates. Good policy is based on good data and models.

6 Summary of Findings

The best regression model found to predict the death rate due to heart disease with these data was the random forest model. It provides the lowest MSE value of 0.02992 while still explaining more than 80% of the variation in the model, and demonstrated a reasonable balance between bias and flexibility. To decrease prediction error, we found we had to sacrifice some of the power of the model: for example, the best multiple linear regression model explained over 90% of the variance, however, because we are handling medical data, accurate predictions are of the utmost importance. All three models demonstrate the importance of including the number of cases of heart disease as a predictor of the number of deaths. After controlling for heart disease cases, the next more significant factor was found to be hospital ranking, then the number of hospitals in a region. Demographic variable associations with heart disease deaths were weak or absent, which was surprising to the researchers and would be a good opportunity for further research. Understanding why demographic variables were not significant, which was an assumption of the research team, would be pressing for future medical studies, as well as collecting complete data, exploring other model diseases, and employing missing data handling methods (such as multiple imputation) for an accurate survey of the state of public health in California.

For the classification model we found the cross validated boosting model with a shrinkage parameter to efficiently predict health professional service area status compared to other models. Other models including pruned and unpruned decision trees and reduced and full logistic regression performed well but the cross validated boosting model with a shrinkage parameter found a balance between having high accuracy 0.877551 and low mean prediction 0.122449 error rate compared to the logistic regression and decision tree models. All six model specifications reported hpsa score and hpsa designation population as important determinants of hpsa status with the final boosting model including metropolitan indicator as key in determining hpsa status. Demographic characteristics of the counties in which the hpsa's are located were initially considered. However after data cleaning didn't fit into our model specifications and were subsequently discarded.

7 Future Directions

The study reveals that the number of patients with heart disease, hospital ranking, and hospital number in a region are significant predictors of mortality rates. The classification model also reveals the importance of hpsa score, hpsa designation population, and metropolitan indicator in predicting health professional service area status. Despite the high accuracy of the models, they lack the necessary demographic characteristics for generalization due to low sample size and data collection methods. The researchers plan to explore further data transformation approaches and combine the predictions of both models into a stacking model to improve their performance. This could enhance the ability to determine mortality rates or hpsa status.

8 Appendix

8.1 Group Member Accomplishments

Roland Abi Developed the research question for the classification, and found the data set for `california health professional shortage area` and cleaned and transformed the data. Completed logistic regression model (full and reduced), as well as cross validation approaches contributed to completing decision tree and boosting models, carried out 627 tasks for model selection and two wrote the classification section of the summary as well as future direction.

Elise Buellesbach

Found and proposed the original CA data set (`California Hospital Inpatient Mortality Rates and Quality Ratings`). Completed the linear regression models (both full and reduced). Wrote the executive summary, summary/conclusion, ethical analysis, and references.

Amanda Concepcion

Spencer Grewe

Acquired CA demographic data (`data4`). Contributed toward cleaning, tidying and joining data sets 1, 3, and 4 for the regression tasks. Carried out 627 tasks involving model selection for regression.

8.2 Statement of Data Availability

The source code, data, and full supporting information are available at the shared repository, which will be made public on the date of the submission of this report. We declare that all authors equally own the contents of the repository.

9 References

10 Supplemental Information

10.1 Appendix Graphics

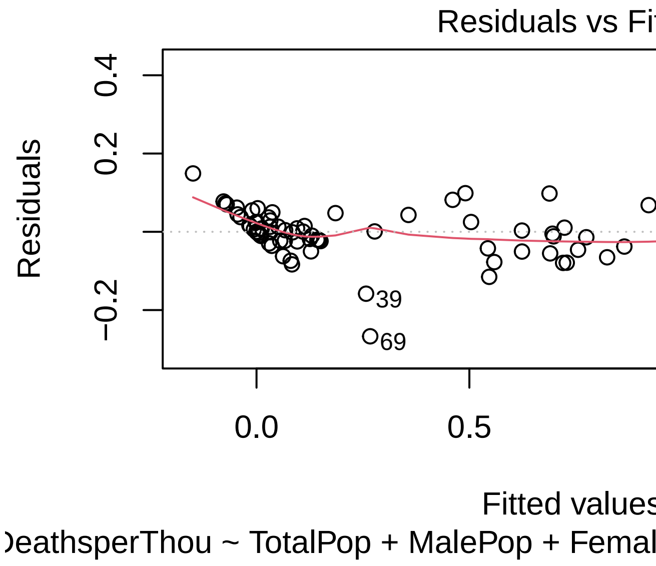
Linear Regression Model ::: {.cell} ::: {.cell-output .cell-output-stderr}

```
Rows: 99 Columns: 40
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): County, Rating
dbl (38): TotalPop, MalePop, FemalePop, MedianAge, MaleMedianAge, FemaleMedi...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

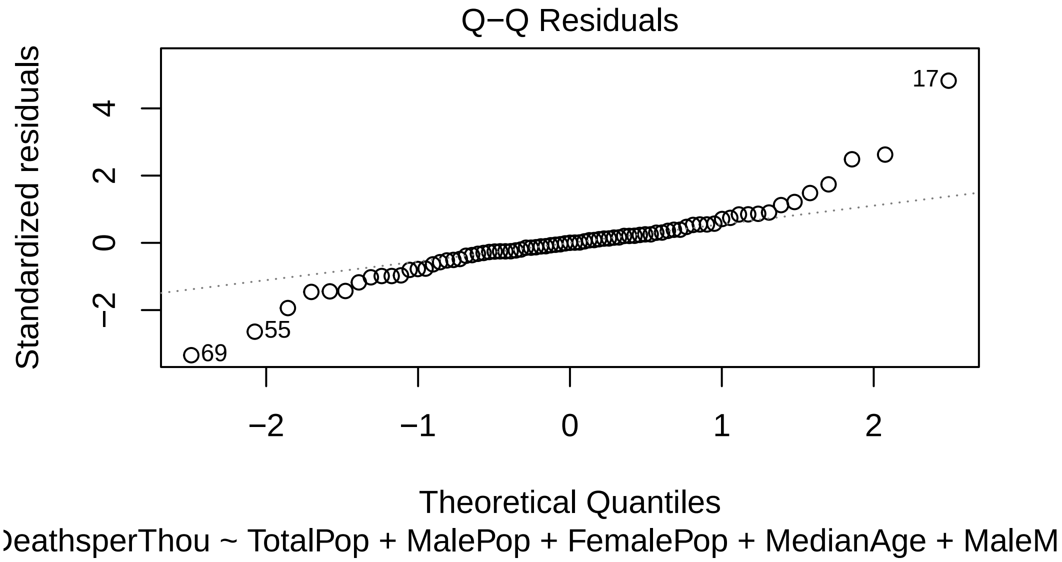
::: :::

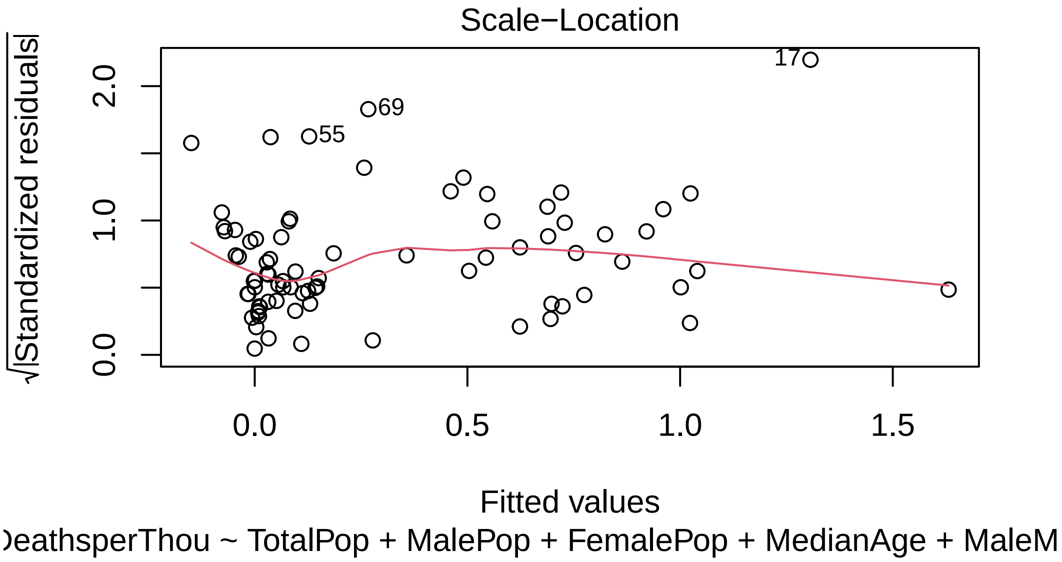
NA Counts for Regression Data ::: {.cell} ::: {.cell-output .cell-output-stdout}

```
         County        TotalPop         MalePop       FemalePop       MedianAge 
              0               0               0               0               0 
  MaleMedianAge FemaleMedianAge     Under5yoPop    Under18yoPop   Pop21andOlder 
              0               0               0               0               0 
  Pop55andOlder   Pop60andOlder   Pop65andOlder      PercFemale   PercUnder18yo 
              0               0               0              12              12 
      num_desig     amer_indian           black        hispanic           asian 
              0               0               0               0               0 
     multi_race      pac_island           white         poverty     labor_force 
              0               0               0               0               0 
     unemployed   median_income     TotalDeaths      TotalCases       prop_hpsa 
              0               0               0               0               0 
  nonwhite_prop          Rating          Deaths   DeathsperThou     unempl_prop 
              0               0               0               0               0 
   poverty_prop     prop55older           Cases      HospNumber    CasesperThou 
              0               0               0               0               0 
    female_prop 
              0 
```

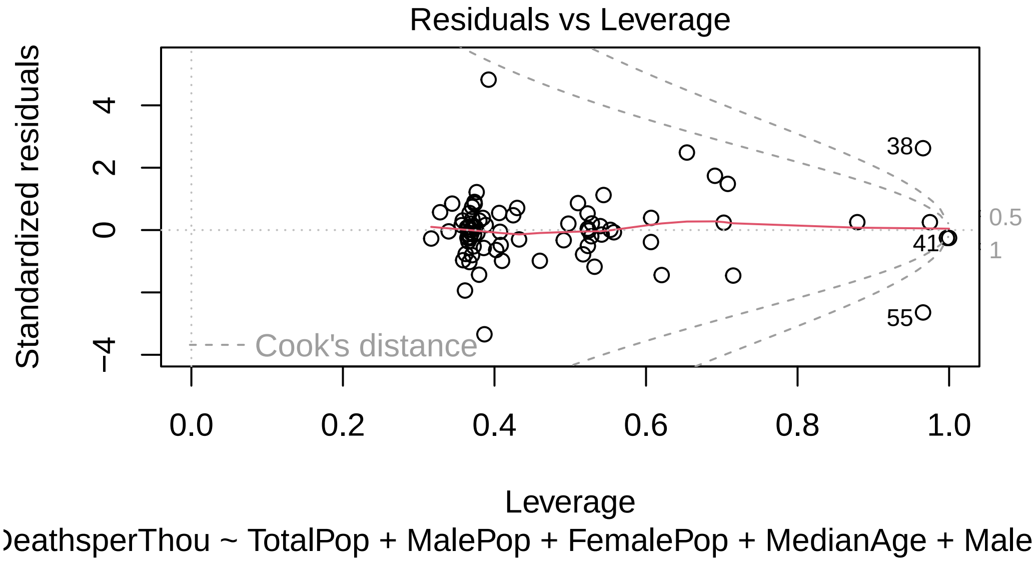


::: ::: Plotting Full Model ::: {.cell} ::: {.cell-output-display} :::





```
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



::: These plots give some insight to the shape of the model. For example, the plotting on the QQ-plot does not demonstrate a strong linear relationship. The edges of the plot deviate from the line suggesting that the model is not perfectly linear. Further, the scattered dots on the Leverage plot show several points on the line of Cook's Distance signifying that they have overstated influence on the model. With these concerns in mind, I want to investigate the linearity of the model further. Does a linear model make sense? To get a better understanding of the data, I created some basic plots to demonstrate the shape of the data and get an idea of if a linear model will make sense here (source on code)

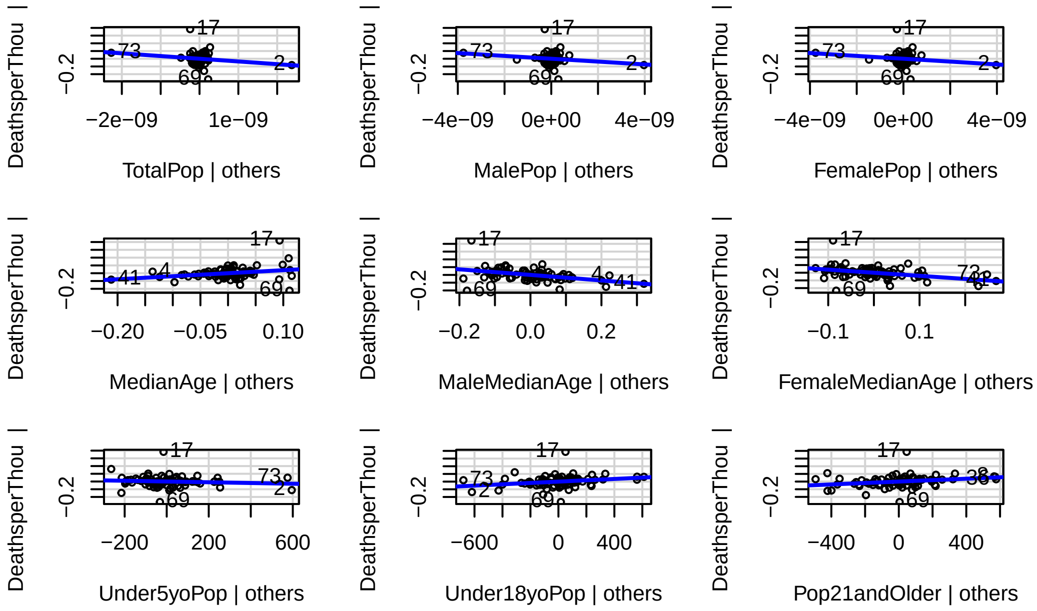
Linear Fit of Variables ::: {.cell} ::: {.cell-output .cell-output-stderr}

```
Loading required package: carData
```

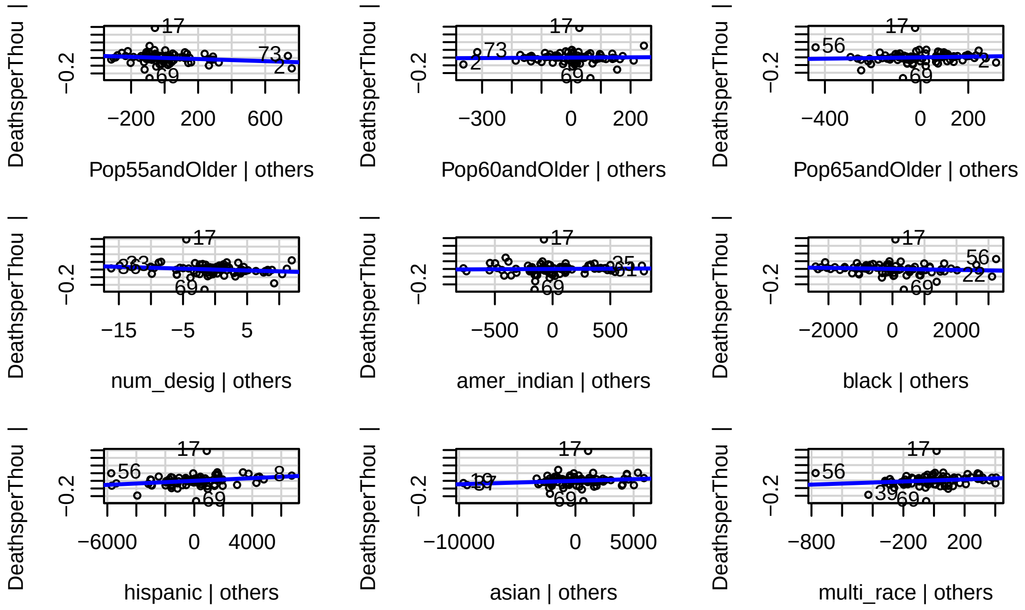
:::

```
Attaching package: 'car'


The following object is masked from 'package:boot':

    logit


The following object is masked from 'package:dplyr':

    recode


The following object is masked from 'package:purrr':

    some


Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
```

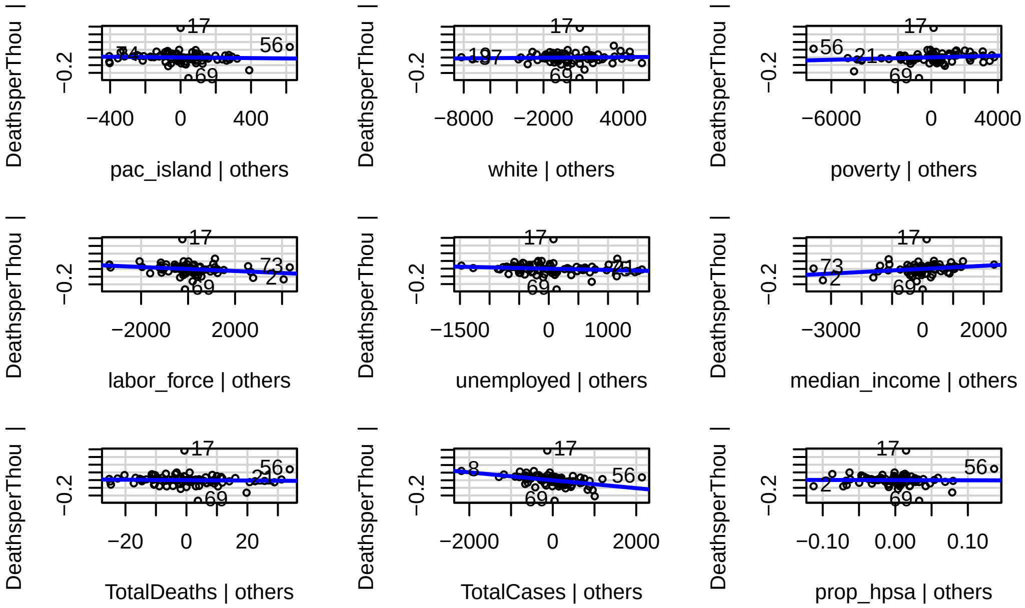
```
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
```



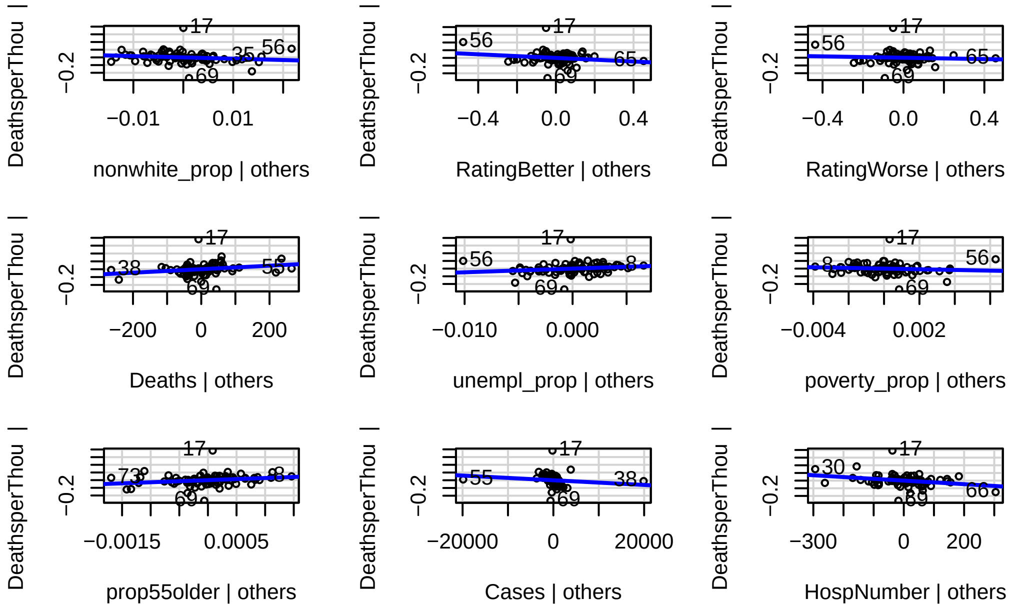
```
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
```



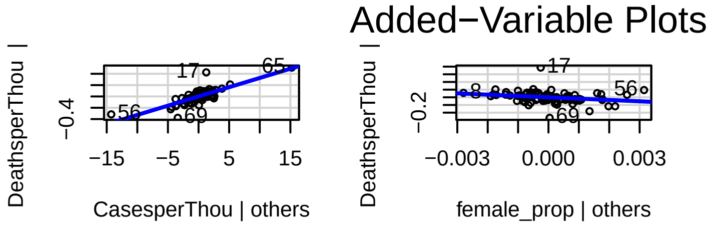
```
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
```



```
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
```



```
Warning in lsfit(mod.mat[, -var], cbind(mod.mat[, var], response), wt = wt, :
'X' matrix was collinear
```



::: Here, we can see that some variables provide a better linear fit than others. For example, variables such as `MaleMedianAge` and `FemaleMedianAge` both are more or less scattered along the regression line. However, other variables such as `MalePop` and `FemalePop` do not fit the line well. Instead, the data is gathered in one spot and has almost no linear shape. These graphs give some reason for concern that the basic full model will not be the strongest. Not all of the variables are linear so a reduced model might be a better fit.

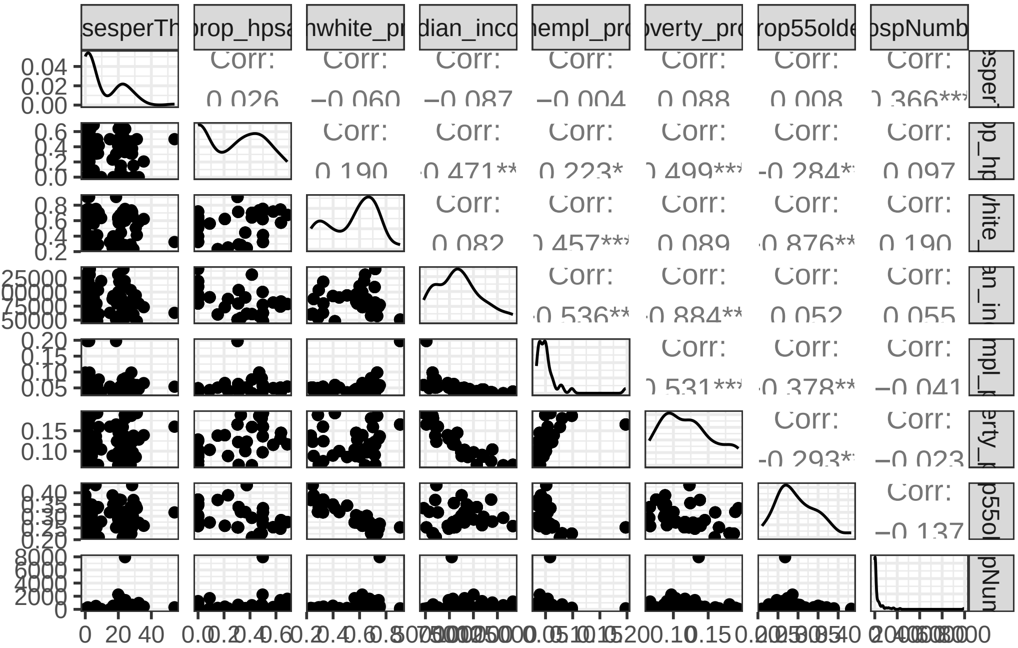


Figure 8: Pairs plot of predictors selected for regression.

Table 6: Coefficients and estimates obtained after LASSO regression at two different L1 penalty strengths.

	Min	X1se
(Intercept)	-0.0936	0.0567
RatingBetter	0.0000	0.0000
RatingWorse	0.0000	0.0000
CasesperThou	0.0328	0.0271
prop_hpsa	0.0000	0.0000
nonwhite_prop	0.0000	0.0000
median_income	0.0000	0.0000
unempl_prop	0.0000	0.0000
poverty_prop	0.0000	0.0000
prop55older	0.4513	0.0000

Table 6: Coefficients and estimates obtained after LASSO regression at two different L1 penalty strengths.

	Min	X1se
HospNumber	-0.0001	0.0000
RatingBetter:CasesperThou	-0.0185	0.0000
RatingWorse:CasesperThou	0.0160	0.0000
RatingBetter:prop55older	-0.0587	0.0000
RatingWorse:prop55older	0.0000	0.0000
RatingBetter:HospNumber	0.0000	0.0000
RatingWorse:HospNumber	0.0000	0.0000

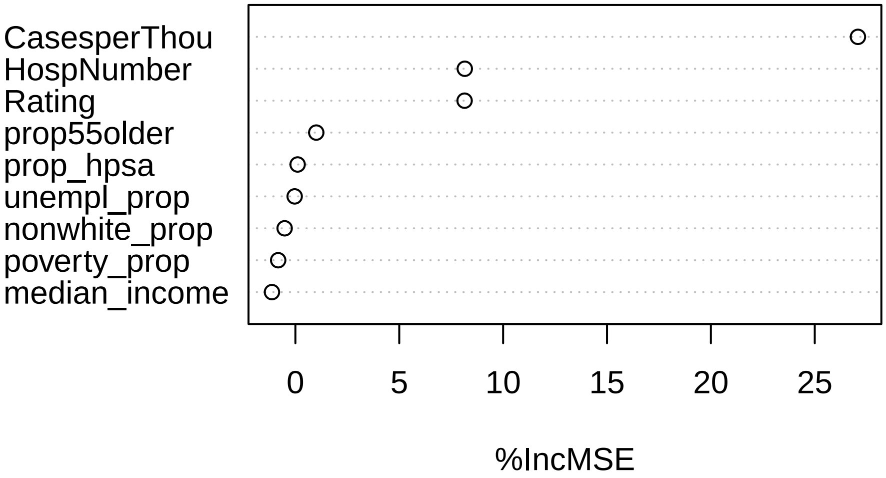


Figure 9: Importance plot of all predictors in the random forest model.