Modeling County-Level Mortality Rates

Introduction

We know that mortality rates from different causes of death vary significantly across regions of the United States and even between counties in the same state, but we don't know a ton about why they vary the way they do. In this project, we wanted to explore what other factors about counties are related to county-level mortality rates from various causes to better understand what some causes of that variation in different causes of death may be. Our analysis and visualizations display actual variation in causes of death in 2014 between counties and how well our multiple models of mortality rates do in predicting the actual mortality rates, showing both the power and the shortcomings of using statistical modeling techniques to estimate causes of death based on other factors about counties.

Motivation/Purpose

It is understood that mortality rates from different causes vary between regions of the United States and even between counties in the same state, but those differences and potential reasons for them have not been systematically explored. Our hope in conducting this project is to better understand some of those differences and to be able to visually present them, as well as the results of some statistical models we made to predict them.

There are several audiences this project could serve different purposes for. One is public health officials - it could be very helpful for them to have a better grasp on actual mortality rates in the areas they operate in, and what factors have been found to be determinants of those mortality rates. A second potential purpose is informing policymakers and government officials of some of the health impacts of decisions they make. While not every variable we consider is completely or, in some cases, even remotely controlled by government action, it could be helpful to understand the impact on mortality rates of, say, the percentage of people who are uninsured or the child poverty rate. Finally, this project will hopefully be part of a conversation with other researchers in the literature on causes of death and their determinants.

Background

In general, little background information is required to understand our project except the knowledge that medical professionals usually record a broad category of cause of death when someone dies. Those records, and the estimates created from them by previous researchers, form the basis of our project. To understand our data wrangling, it is necessary to know that FIPS, or Federal Information Processing Standard, codes uniquely identify US counties. Besides that, the other variables we use are either defined as they are entered or are common terms explained in the variable names.

Data Sources

The first data source we used was a compilation of county-level data from a variety of sources compiled by Github user Deleetdk. Most of the data sources included are those from a paper published by Emil O. W. Kirkegaard in the journal *Open Quantitative Sociology and Political Science* in 2016, "Inequality across US counties: an S factor analysis". This data source was, very luckily for us, an RData file on GitHub that was very easy to download.

The demographic data (such as racial composition, gender, and age) come from the United States Census Bureau's annual American Community Survey through the American FactFinder. This is also the source for some economic indicators (educational attainment, sector composition of employment). Another data source is the American Human Development Index of the Measure of America, which contains health, education, and income indicators. Many health indicators come from the Robert Wood Johnson Foundation's County Health Rankings & Roadmaps. Finally, the county-level electoral data comes from the failing New York Times.

The first step of data wrangling with this dataset was selecting the variables we wanted from it. We then removed the state-level rows, which all have FIPS codes of X000, so that the data was only counties.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
load("USA_county_data.RData")
countydata <- USA_county_data %>% select(fips, name_16, rep16_frac,
    dem16_frac, rep16_frac2, dem16_frac2, statecode_prev, rep12_frac,
    dem12_frac, rep12_frac2, dem12_frac2, rep08_frac, dem08_frac,
    rep08_frac2, dem08_frac2, Less.Than.High.School, At.Least.High.School.Diploma,
    At.Least.Bachelor.s.Degree, Graduate.Degree, School.Enrollment,
    Median.Earnings.2010.dollars, White.not.Latino.Population,
    African.American.Population, Native.American.Population,
    Asian.American.Population, Latino.Population, Children.Under.6.Living.in.Poverty,
    Adults.65.and.Older.Living.in.Poverty, Total.Population,
    Preschool.Enrollment.Ratio.enrolled.ages.3.and.4, Poverty.Rate.below.federal.poverty.threshold,
    Child.Poverty.living.in.families.below.the.poverty.line,
    Management.professional.and.related.occupations, Service.occupations,
    Sales.and.office.occupations, Farming.fishing.and.forestry.occupations,
    Construction.extraction.maintenance.and.repair.occupations,
    Production.transportation.and.material.moving.occupations,
    State, median_age, Poor.physical.health.days, Poor.mental.health.days,
    Low.birthweight, Teen.births, Children.in.single.parent.households,
    Adult.smoking, Adult.obesity, Diabetes, Sexually.transmitted.infections,
    HIV.prevalence.rate, Uninsured, Unemployment)

countydata <- countydata %>% filter(!fips %in% seq(0, 1e+05,
    by = 1000)) %>% arrange(fips)
```

The key dataset for this analysis is a set of estimates of "annual mortality rates by US county for 21

mutually exclusive causes of death from 1980 through 2014" constructed by Laura Dwyer-Lindgren, Amelia Bertozzi-Villa, Rebecca Stubbs, et al. from the National Center for Health Statistics' National Vital Statistics System and published in the Journal of the American Medical Association. We downloaded the dataset itself from Kaggle, which it was provided to by the Institute for Health Metrics and Evaluation.

The causes of death (and our short names for them) are:

HIV: HIV-AIDS and Tuberculosis

INF: Diarrhea, lower respiratory, and other common infectious diseases

TROP: Neglected tropical diseases and malaria

MAT: Maternal disorders

NEON: Neonatal disorders

NUT: Nutritional deficiencies

OTHC: Other communicable, maternal, neonatal, and nutritional diseases

NEOP: Neoplasms

CHRON: Chronic respiratory diseases

CIRR: Cirrhosis and other chronic liver diseases

DIG: Digestive diseases

NEUR: Neurological disorders

MENSUB: Mental and substance use disorders

DIAB: Diabetes, urogenital, blood, and endocrine diseases

MUSC: Musculoskeletal disorders

OTHN: Other non-communicable diseases

TRAN: Transport injuries

UNIN: Unintentional injuries

SELF: Self-harm and interpersonal violence

WAR: Forces of nature, war, and legal intervention

This dataset was an Excel file with estimates with estimated mortality rates for each county for each of 20 causes of mortality for each of many years (1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2014). The first thing we had to do was change each estimate from a midpoint and a confidence interval to just the midpoint. We also renamed the sheets and columns to make them easier to work with. Then, because each cause of death was its own sheet, we had to make a function to read in all of the sheets to one dataframe in RStudio.

```
library(readxl)
read_excel_allsheets <- function(filename) {
    sheets <- readxl::excel_sheets(filename)
    x <- lapply(sheets, function(X) readxl::read_excel(filename,
        sheet = X))
    names(x) <- sheets
    x
}
mortalityrates <- read_excel_allsheets("MortalityRatesClean.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
mortalityrates <- as.data.frame(mortalityrates)
```

Because each cause of death was its own sheet, when we imported the data there were 20 columns of the FIPS codes and 20 columns of location (the county names). As a result, we selected one of the location and FIPS and all of the mortality rate estimates to keep.

```
require(dplyr)
mortalityrates <- rename(mortalityrates, fips = HIV.FIPS)
mortalityrates <- rename(mortalityrates, location = HIV.Location)
mortalityrates <- mortalityrates %>% select(location, fips, HIV.HIV80,
    HIV.HIV85, HIV.HIV90, HIV.HIV95, HIV.HIV00, HIV.HIV05, HIV.HIV10,
    HIV.HIV14, INF.INF80, INF.INF85, INF.INF90, INF.INF95, INF.INF00,
    INF.INF05, INF.INF10, INF.INF14, TROP.TROP80, TROP.TROP85,
    TROP.TROP90, TROP.TROP95, TROP.TROP00, TROP.TROP05, TROP.TROP10,
    TROP.TROP14, MAT.MAT80, MAT.MAT85, MAT.MAT90, MAT.MAT95,
    MAT.MAT00, MAT.MAT05, MAT.MAT10, MAT.MAT14, NEON.NEON80,
    NEON.NEON85, NEON.NEON90, NEON.NEON95, NEON.NEON00, NEON.NEON05,
    NEON.NEON10, NEON.NEON14, NUT.NUT80, NUT.NUT85, NUT.NUT90,
    NUT.NUT95, NUT.NUT00, NUT.NUT05, NUT.NUT10, NUT.NUT14, OTHC.OTHC80,
    OTHC.OTHC85, OTHC.OTHC90, OTHC.OTHC95, OTHC.OTHC00, OTHC.OTHC05,
    OTHC.OTHC10, OTHC.OTHC14, NEOP.NEOP80, NEOP.NEOP85, NEOP.NEOP90,
    NEOP.NEOP95, NEOP.NEOP00, NEOP.NEOP05, NEOP.NEOP10, NEOP.NEOP14,
    CHRON.CHRON80, CHRON.CHRON85, CHRON.CHRON90, CHRON.CHRON95,
    CHRON.CHRON00, CHRON.CHRON05, CHRON.CHRON10, CHRON.CHRON14,
    CIRR.CIRR80, CIRR.CIRR85, CIRR.CIRR90, CIRR.CIRR95, CIRR.CIRR00,
    CIRR.CIRR05, CIRR.CIRR10, CIRR.CIRR14, DIG.DIG80, DIG.DIG85,
    DIG.DIG90, DIG.DIG95, DIG.DIG00, DIG.DIG05, DIG.DIG10, DIG.DIG14,
    NEUR.NEUR80, NEUR.NEUR85, NEUR.NEUR90, NEUR.NEUR95, NEUR.NEUR00,
    NEUR.NEUR05, NEUR.NEUR10, NEUR.NEUR14, MENSUB.MENSUB80, MENSUB.MENSUB85,
    MENSUB.MENSUB90, MENSUB.MENSUB95, MENSUB.MENSUB00, MENSUB.MENSUB05,
    MENSUB.MENSUB10, MENSUB.MENSUB14, DIAB.DIAB80, DIAB.DIAB85,
    DIAB.DIAB90, DIAB.DIAB95, DIAB.DIAB00, DIAB.DIAB05, DIAB.DIAB10,
    DIAB.DIAB14, MUSC.MUSC80, MUSC.MUSC85, MUSC.MUSC90, MUSC.MUSC95,
    MUSC.MUSC00, MUSC.MUSC05, MUSC.MUSC10, MUSC.MUSC14, OTHN.OTHN80,
    OTHN.OTHN85, OTHN.OTHN90, OTHN.OTHN95, OTHN.OTHN00, OTHN.OTHN05,
    OTHN.OTHN10, OTHN.OTHN14, TRAN.TRAN80, TRAN.TRAN85, TRAN.TRAN90,
    TRAN.TRAN95, TRAN.TRAN00, TRAN.TRAN05, TRAN.TRAN10, TRAN.TRAN14,
    UNIN.UNIN80, UNIN.UNIN85, UNIN.UNIN90, UNIN.UNIN95, UNIN.UNIN00,
    UNIN.UNIN05, UNIN.UNIN10, UNIN.UNIN14, SELF.SELF80, SELF.SELF85,
    SELF.SELF90, SELF.SELF95, SELF.SELF00, SELF.SELF05, SELF.SELF10,
    SELF.SELF14, WAR.WAR80, WAR.WAR85, WAR.WAR90, WAR.WAR95,
    WAR.WAR00, WAR.WAR05, WAR.WAR10, WAR.WAR14)
```

Similar to with the other dataset, there were state rows as well as county rows. However, in this dataset the states had low FIPS codes (below 100) instead of FIPS codes of multiples of 1000. Additionally, the mortality rates dataset had counties in Puerto Rico (FIPS codes in the 72000s), but we removed these because they did not have data in the Deleetdk data.

```
mortalityrates <- mortalityrates %>% filter(fips > 100) %>% filter(fips <
    70000) %>% arrange(fips)
```

At this point, we combined the two datasets.

```
require(dplyr)
countyrates <- inner_join(countydata, mortalityrates, by = c("fips"))
```

When we thought we needed the latitude and longitude of each county, we downloaded a file that contained that information and renamed and selected the variables for them.

```
require(dplyr)
`2015_Gaz_counties_national` <- read.delim("2015_Gaz_counties_national.txt")
`2015_Gaz_counties_national` <- rename(`2015_Gaz_counties_national`,
    fips = GEOID)
`2015_Gaz_counties_national` <- rename(`2015_Gaz_counties_national`,
    latitude = INTPTLAT)
`2015_Gaz_counties_national` <- rename(`2015_Gaz_counties_national`,
    longitude = INTPTLONG)
`2015_Gaz_counties_national` <- `2015_Gaz_counties_national` %>%
    select(fips, latitude, longitude)
```

We then merged that dataset into our main dataframe, countyrates.

```
countyrates <- left_join(countyrates, `2015_Gaz_counties_national`,
    by = c("fips"))
```

After combining the datasets, we noticed there were several observations with missing data for some variables. In particular, all of Alaska was missing most of the Deleetdk data and a few stray observations were missing the education data and a number of other variables. We removed those observations.

```
countyratesnogaps <- countyrates %>% filter(fips > 100) %>% filter(fips <
    70000) %>% filter(!is.na(rep16_frac)) %>% filter(!is.na(At.Least.High.School.Diploma)) %>%
    arrange(fips)
```

We also chose to remove some of the variables that had a lot of missing data by selecting the others.

Finally, we formally created the dataframe we would use of all the county observations with tidy data for all of our variables of interest.

```
countyratesfullcases <- countyratesnogaps[complete.cases(countyratesnogaps),
    ]
```

Model Creation/Statistical Computation

We created two models for each of the twenty causes of death - a regression tree and a k nearest neighbors model.

Regression trees classify counties into regions based on their values of the explanatory variables, then assigns every county in a given region the average of the counties in that region.

k-nearest neighbors look at the k counties, in this case 5 counties, closest to the county of interest in terms of their values of the explanatory variables, and predicts the mortality rate for that county as the average of the mortality rates of those 5 counties for that cause of death.

```
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
require(rpart)
```

```
## Loading required package: rpart
```

```
require(rpart.plot)
```

```
## Loading required package: rpart.plot
```

```
library(tree)
```

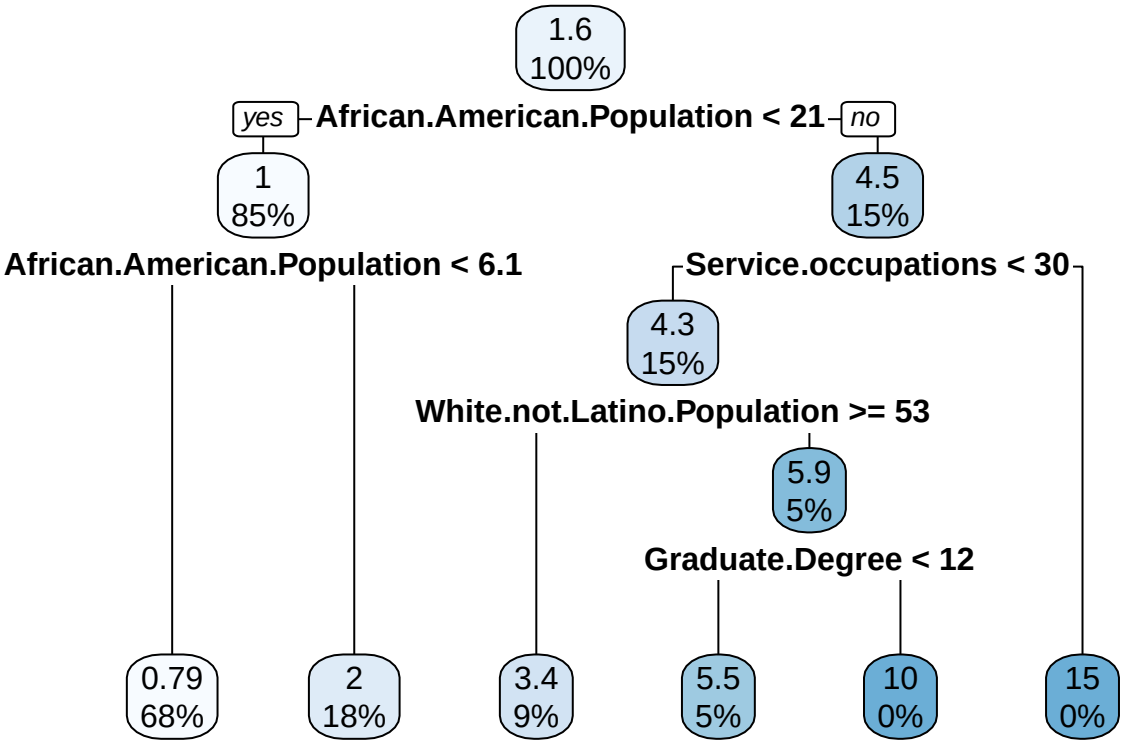
To give a sense for our model-building process, we have included the R code for the model creation for the cause of death, HIV-AIDS and tuberculosis. All of the other models follow the same format, and only the regression trees and the mean square errors from the k-nearest neighbors model are shown for them.

HIV: HIV-AIDS and Tuberculosis

Regression tree

```
fips <- countyratesfullcases$fips
```

```
require(dplyr)
set.seed(1)
fitControl <- trainControl(method = "cv")
tr.HIV14 <- train(HIV.HIV14 ~ rep16_frac + dem16_frac + Less.Than.High.School +
    At.Least.High.School.Diploma + At.Least.Bachelor.s.Degree +
    Graduate.Degree + School.Enrollment + Median.Earnings.2010.dollars +
    White.not.Latino.Population + African.American.Population +
    Native.American.Population + Asian.American.Population +
    Latino.Population + Adults.65.and.Older.Living.in.Poverty +
    Total.Population + Poverty.Rate.below.federal.poverty.threshold +
    Child.Poverty.living.in.families.below.the.poverty.line +
    Management.professional.and.related.occupations + Service.occupations +
    Sales.and.office.occupations + Farming.fishing.and.forestry.occupations +
    Construction.extraction.maintenance.and.repair.occupations +
    Production.transportation.and.material.moving.occupations +
    median_age + Children.in.single.parent.households + Adult.obesity +
    Diabetes + Uninsured + Unemployment, data = countyratesfullcases,
    method = "rpart2", trControl = fitControl, tuneGrid = data.frame(maxdepth = 1:20))
rpart.plot(tr.HIV14$finalModel)
```




```
mortrate <- countyratesfullcases$HIV.HIV14
cause <- rep("HIV", 3110)
cause <- as.data.frame(cause)
model <- rep("real", 3110)
model <- as.data.frame(model)
HIVreal <- cbind(fips, cause, model, mortrate)
```

```
realHIV14 <- countyratesfullcases$HIV.HIV14
preds.tr.HIV14 <- predict(tr.HIV14$finalModel, newdata = countyratesfullcases)
HIV14 <- cbind.data.frame(preds.tr.HIV14, realHIV14)
HIV14 <- HIV14 %>% mutate(HIVtreediff = ((preds.tr.HIV14 - realHIV14)/realHIV14))
mortrate <- HIV14$HIVtreediff
cause <- rep("HIV", 3110)
cause <- as.data.frame(cause)
model <- rep("tree", 3110)
model <- as.data.frame(model)
HIVtree <- cbind(fips, cause, model, mortrate)
```

Regression tree mean square error

```
treeHIV14.mse <- mean((preds.tr.HIV14 - realHIV14)^2)
treeHIV14.mse
```

```
## [1] 2.003797
```

k-Nearest Neighbors mean square error

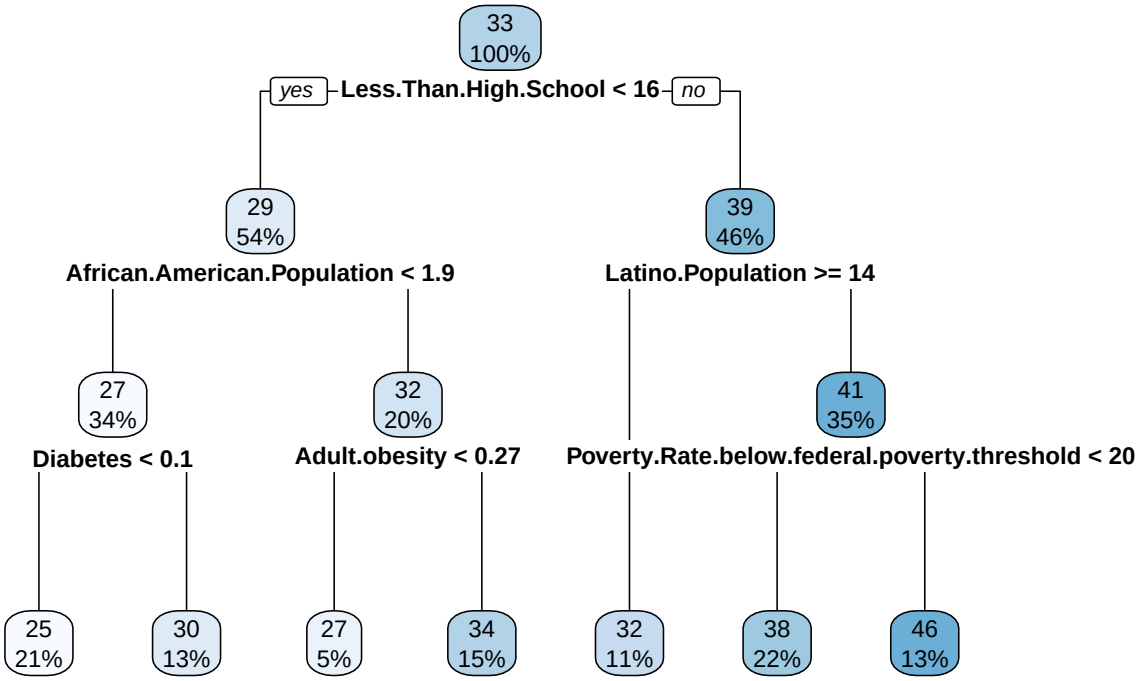
```
set.seed(4747)
fitControl <- trainControl(method = "none")
knn.tr.HIV14 <- train(HIV.HIV14 ~ rep16_frac + dem16_frac + Less.Than.High.School +
    At.Least.High.School.Diploma + At.Least.Bachelor.s.Degree +
    Graduate.Degree + School.Enrollment + Median.Earnings.2010.dollars +
```

```
    White.not.Latino.Population + African.American.Population +
    Native.American.Population + Asian.American.Population +
    Latino.Population + Adults.65.and.Older.Living.in.Poverty +
    Total.Population + Poverty.Rate.below.federal.poverty.threshold +
    Child.Poverty.living.in.families.below.the.poverty.line +
    Management.professional.and.related.occupations + Service.occupations +
    Sales.and.office.occupations + Farming.fishing.and.forestry.occupations +
    Construction.extraction.maintenance.and.repair.occupations +
    Production.transportation.and.material.moving.occupations +
    median_age + Children.in.single.parent.households + Adult.obesity +
    Diabetes + Uninsured + Unemployment, data = countyratesfullcases,
    method = "knn", trControl = fitControl, tuneGrid = data.frame(k = 5))

preds.knn.tr.HIV14 <- predict(knn.tr.HIV14, newdata = countyratesfullcases)
realknnHIV14 <- countyratesfullcases$HIV.HIV14
knnHIV14 <- cbind.data.frame(preds.knn.tr.HIV14, realknnHIV14)
knnHIV14 <- knnHIV14 %>% mutate(HIVknndiff = ((preds.knn.tr.HIV14 -
    realknnHIV14)/realknnHIV14))
mortrate <- knnHIV14$HIVknndiff
cause <- rep("HIV", 3110)
cause <- as.data.frame(cause)
model <- rep("knn", 3110)
model <- as.data.frame(model)
HIVknn <- cbind(fips, cause, model, mortrate)

knnHIV14.mse <- mean((preds.knn.tr.HIV14 - realknnHIV14)^2)
knnHIV14.mse
```

```
## [1] 3.017696
```

INF: Diarrhea, lower respiratory, and other common infectious diseases

Regression tree and mean square error




```
## [1] 66.67221
```

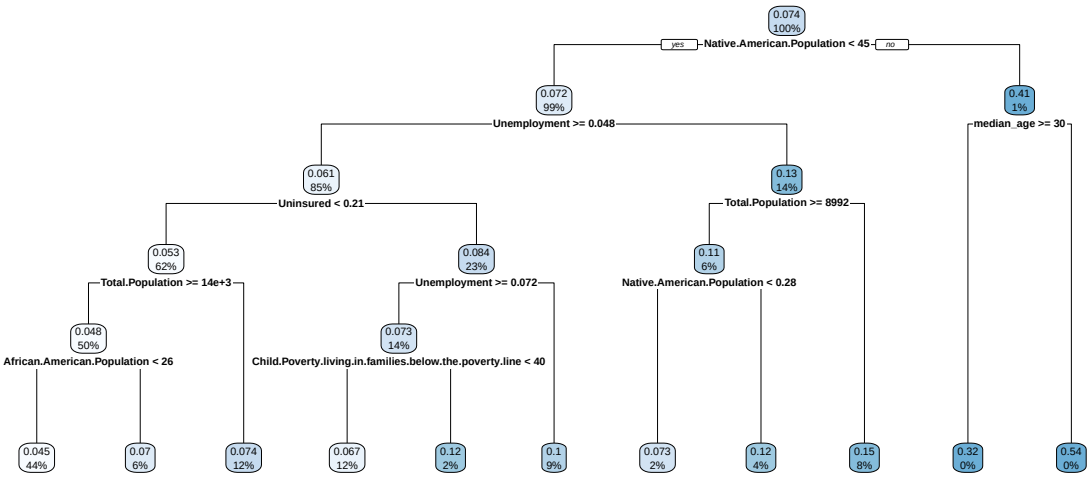
k-Nearest Neighbors mean square error

```
## [1] 79.45923
```

TROP: Neglected tropical diseases and malaria

Regression tree and mean square error

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



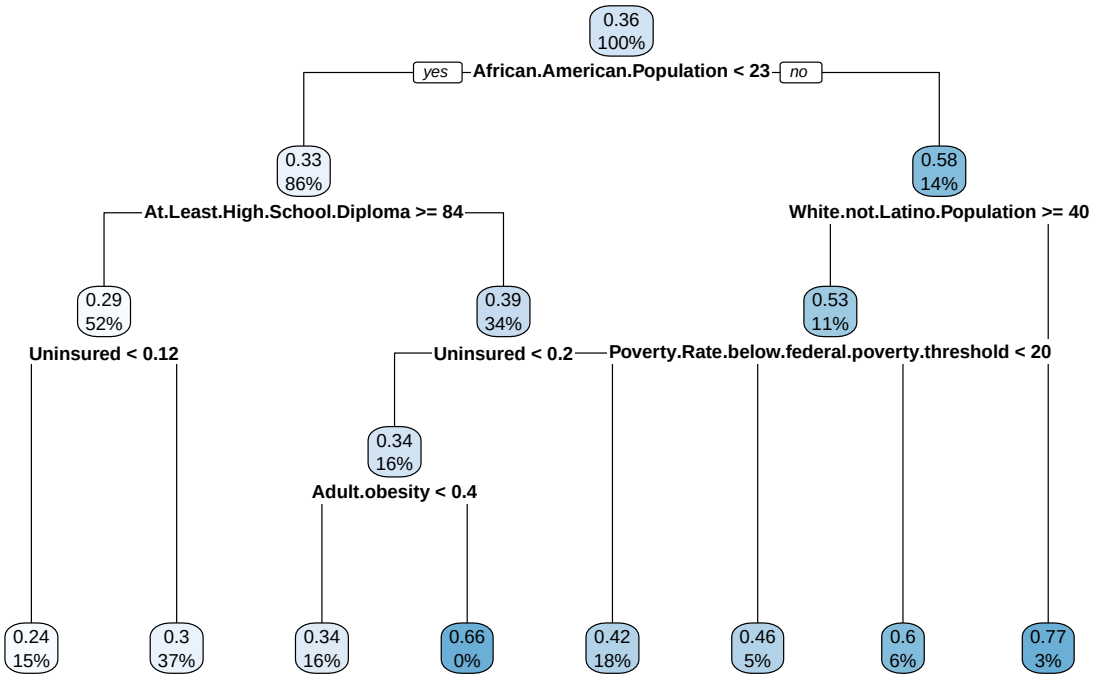

```
## [1] 0.001189699
```

k-Nearest Neighbors mean square error

```
## [1] 0.001668828
```

MAT: Maternal disorders

Regression tree and mean square error



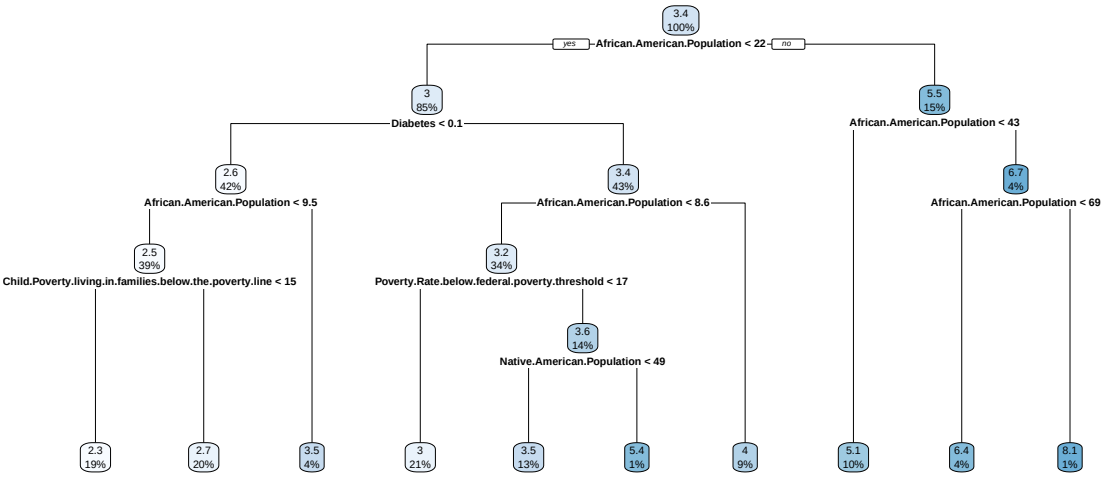

```
## [1] 0.009208227
```

k-Nearest Neighbors mean square error

```
## [1] 0.01464281
```

NEON: Neonatal disorders

Regression tree and mean square error



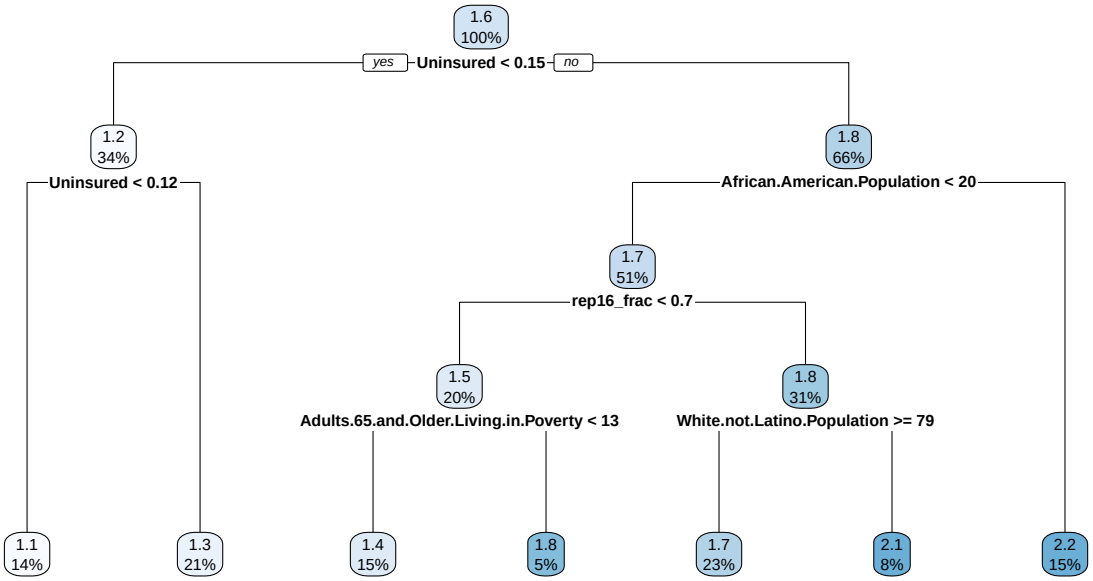

```
## [1] 0.4503517
```

k-Nearest Neighbors mean square error

```
## [1] 1.096217
```

NUT: Nutritional deficiencies

Regression tree and mean square error



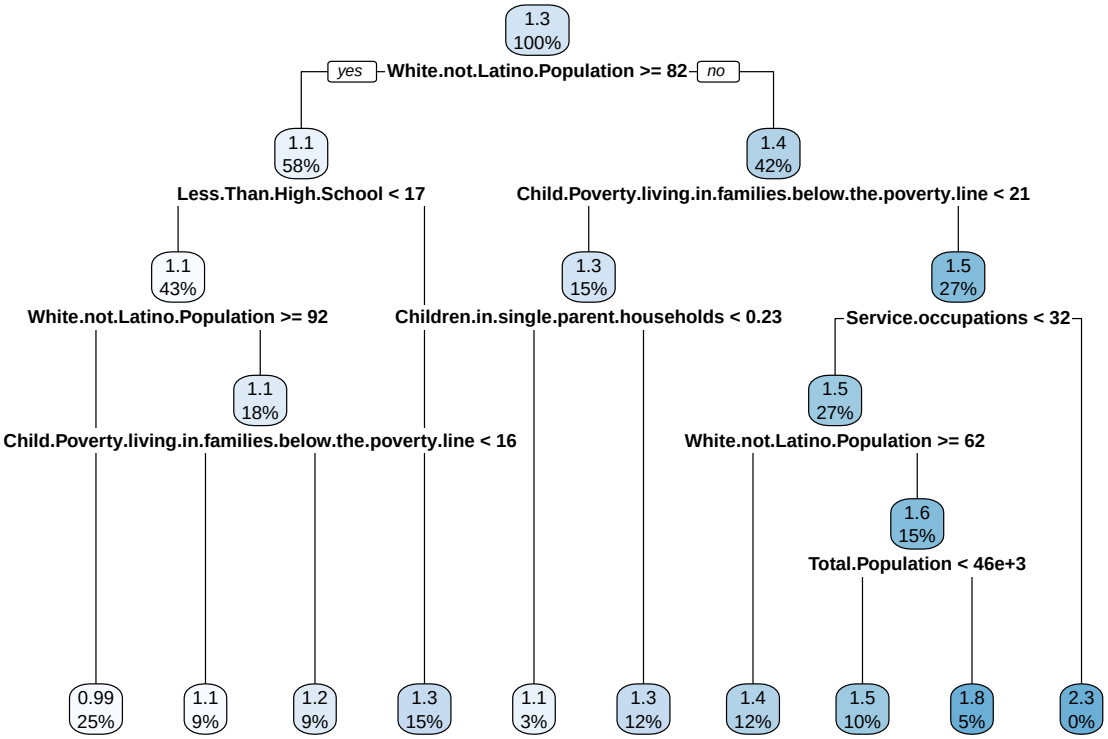

```
## [1] 0.3479154
```

k-Nearest Neighbors mean square error

```
## [1] 0.3380761
```

OTHC: Other communicable, maternal, neonatal, and nutritional diseases

Regression tree and mean square error



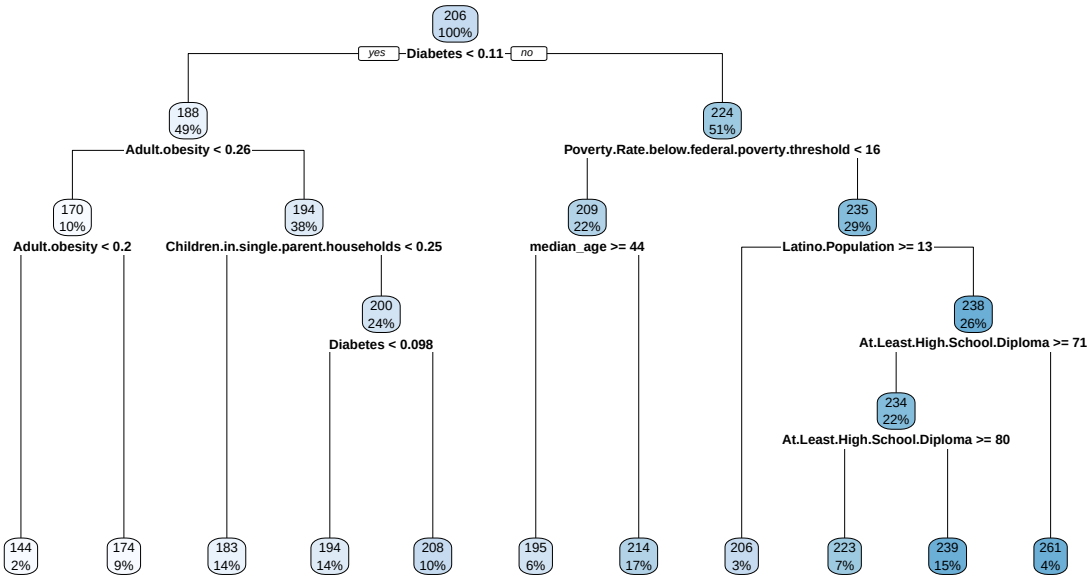

```
## [1] 0.05241569
```

k-Nearest Neighbors mean square error

```
## [1] 0.0618921
```

NEOP: Neoplasms

Regression tree and mean square error



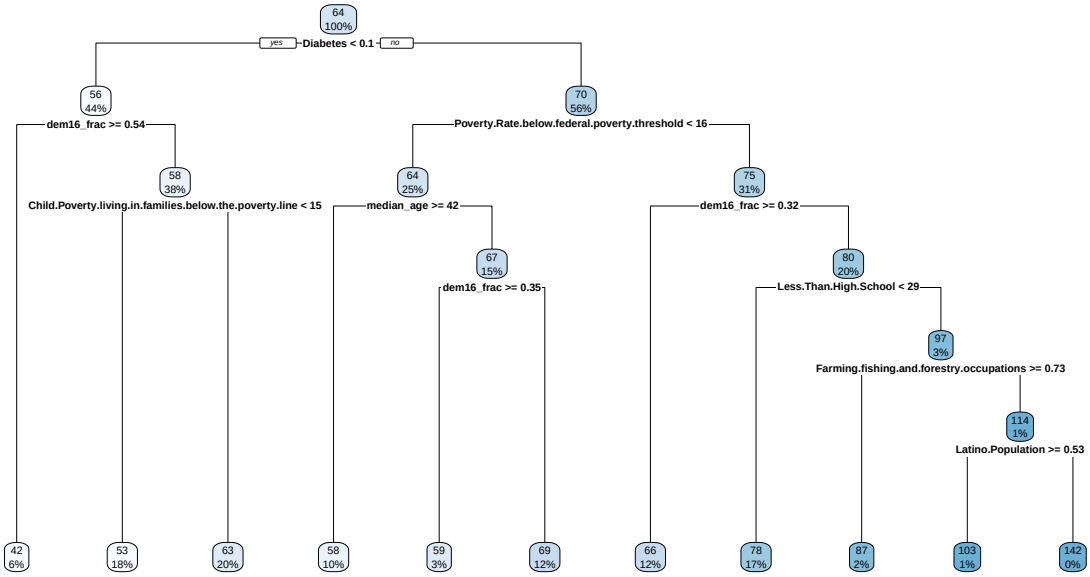

```
## [1] 437.6976
```

k-Nearest Neighbors mean square error

```
## [1] 622.1279
```

CHRON: Chronic respiratory diseases

Regression tree and mean square error




```
## [1] 158.0894
```

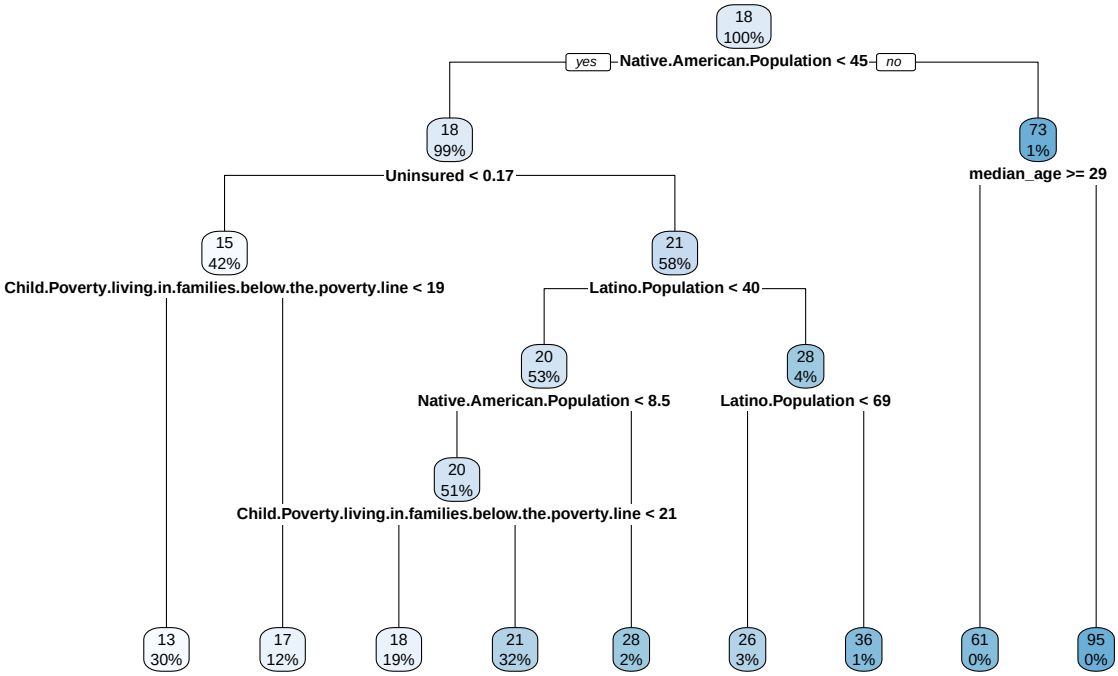
k-Nearest Neighbors mean square error

```
## [1] 189.6379
```

CIRR: Cirrhosis and other chronic liver diseases

Regression tree and mean square error

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



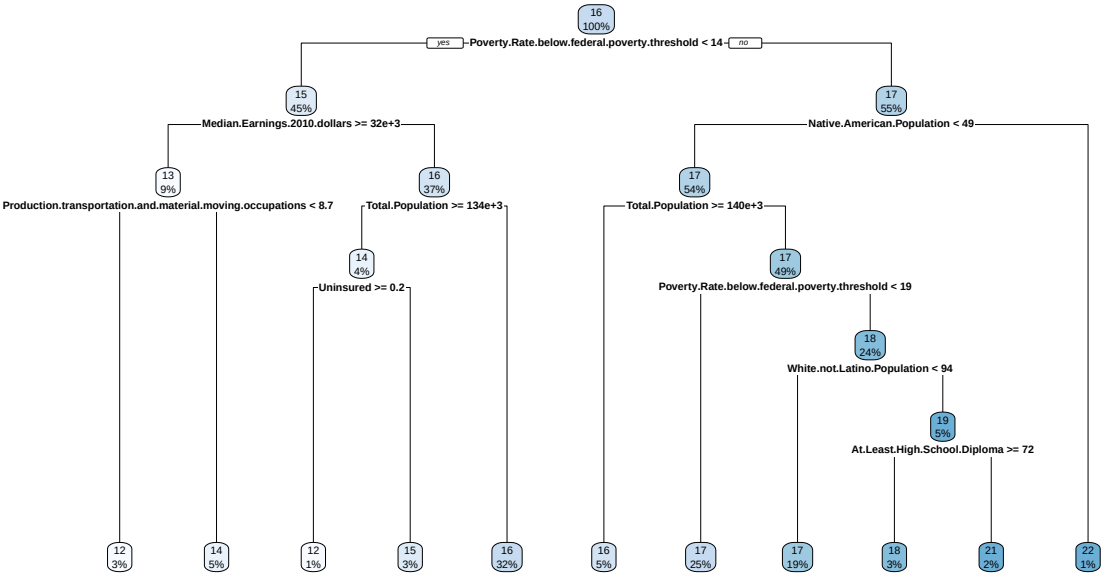

```
## [1] 23.49134
```

k-Nearest Neighbors mean square error

```
## [1] 39.78413
```

DIG: Digestive diseases

Regression tree and mean square error



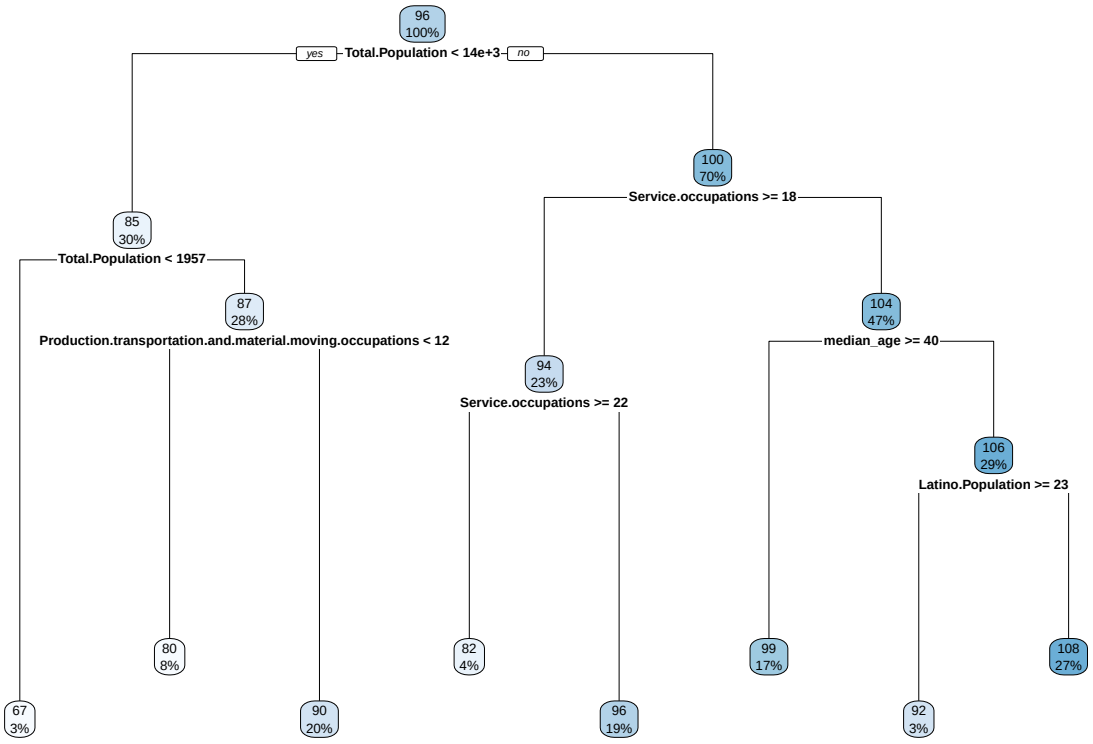

```
## [1] 4.024479
```

k-Nearest Neighbors mean square error

```
## [1] 3.739169
```

NEUR: Neurological disorders

Regression tree and mean square error



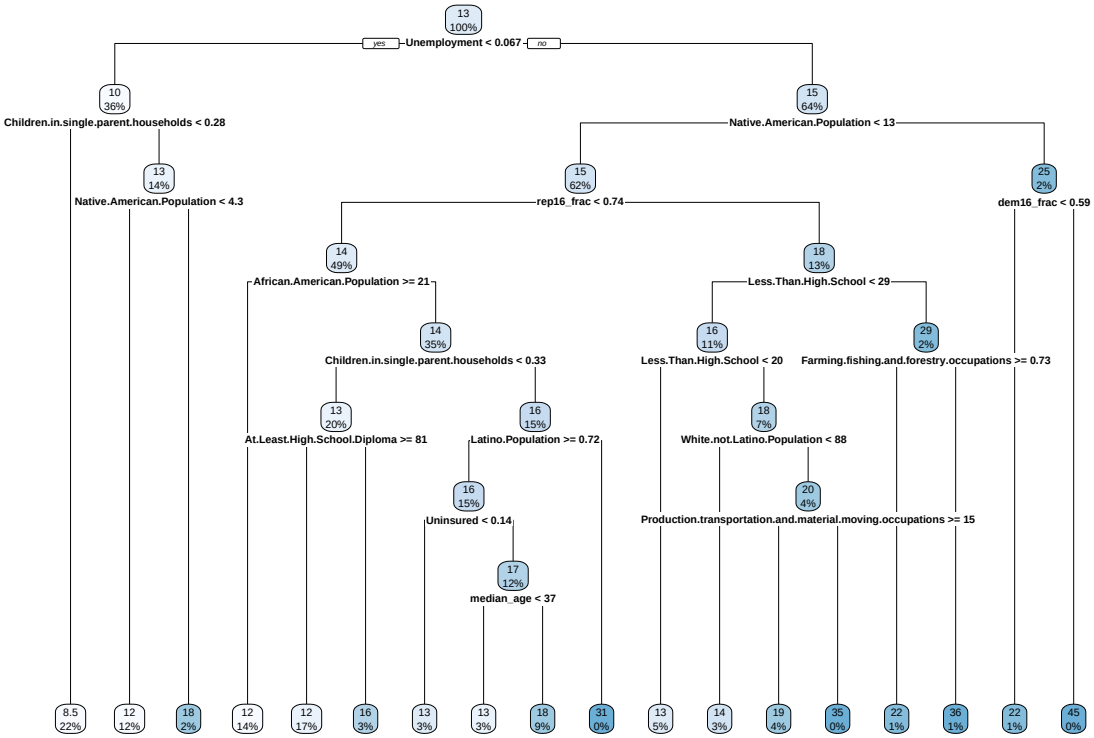

```
## [1] 376.5733
```

k-Nearest Neighbors mean square error

```
## [1] 320.2459
```

MENSUB: Mental and substance use disorders

Regression tree and mean square error



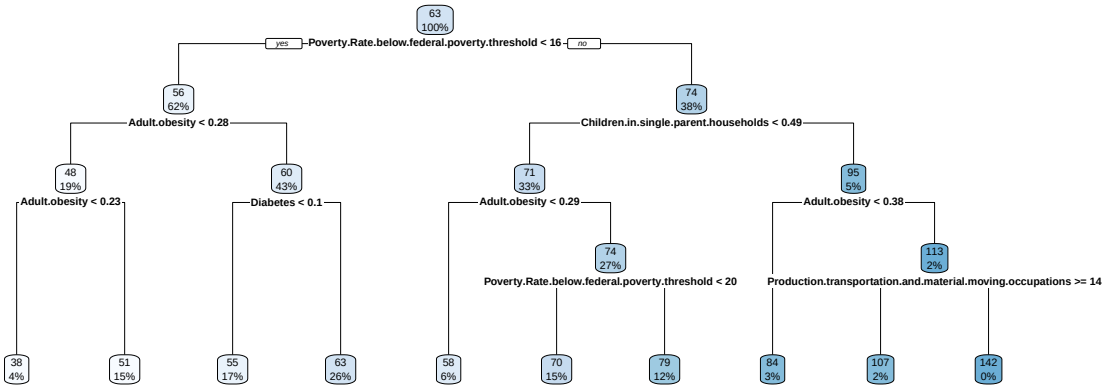

```
## [1] 24.66439
```

k-Nearest Neighbors mean square error

```
## [1] 31.55449
```

DIAB: Diabetes, urogenital, blood, and endocrine diseases

Regression tree and mean square error



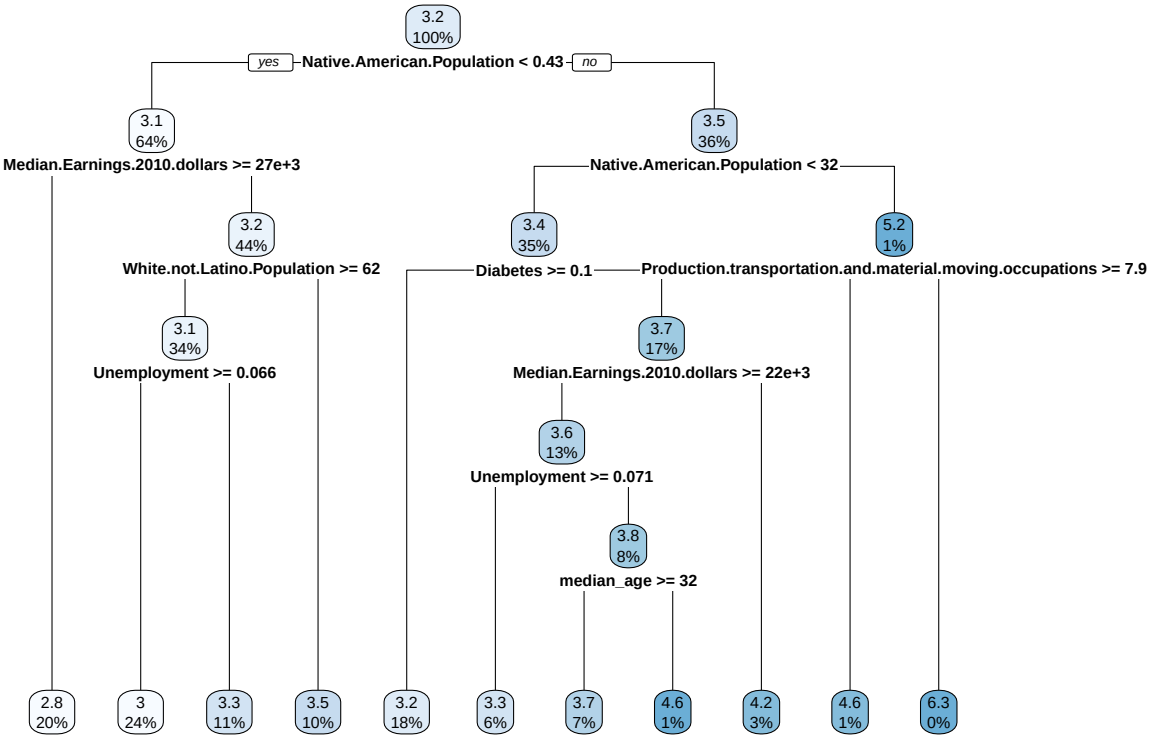

```
## [1] 157.4018
```

k-Nearest Neighbors mean square error

```
## [1] 212.7531
```

MUSC: Musculoskeletal disorders

Regression tree and mean square error



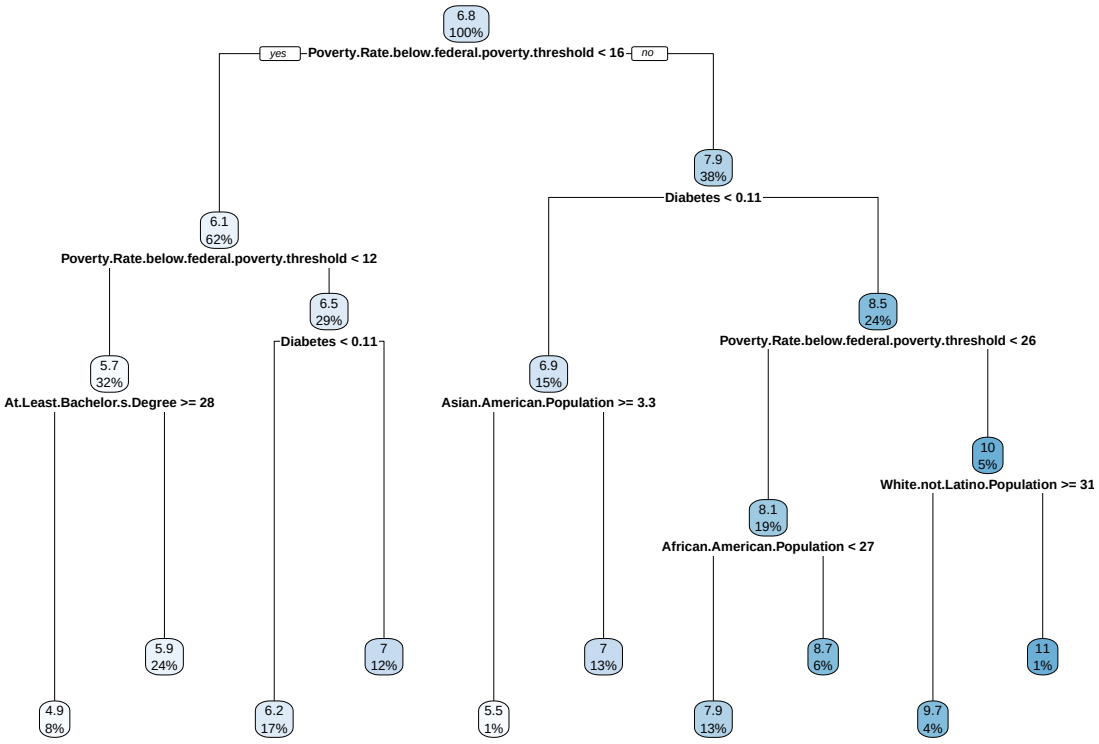

```
## [1] 0.4045811
```

k-Nearest Neighbors mean square error

```
## [1] 0.4037828
```

OTHN: Other non-communicable diseases

Regression tree and mean square error



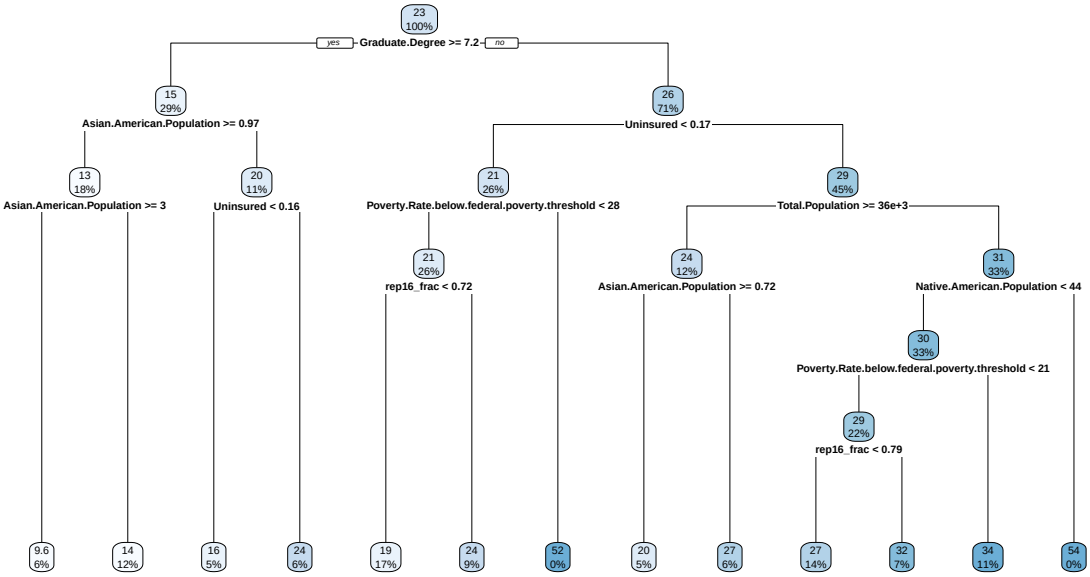

```
## [1] 0.7641577
```

k-Nearest Neighbors mean square error

```
## [1] 1.235623
```

TRAN: Transport injuries

Regression tree and mean square error




```
## [1] 31.85381
```

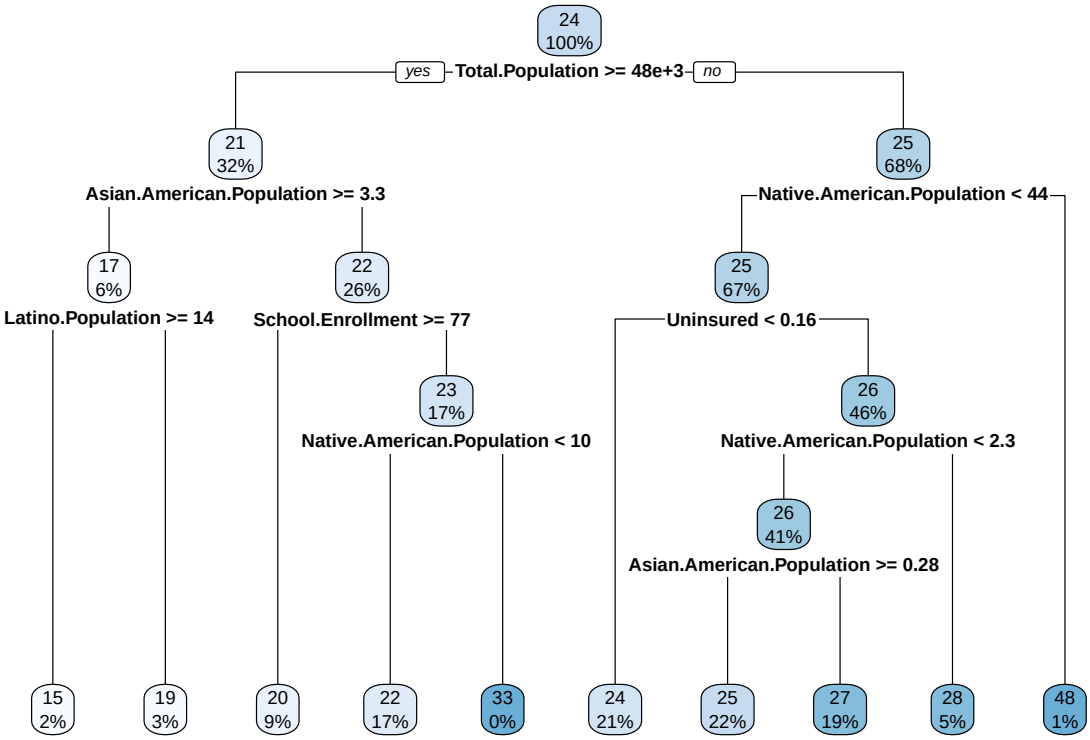
k-Nearest Neighbors mean square error

```
## [1] 37.92881
```

UNIN: Unintentional injuries

Regression tree and mean square error

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```



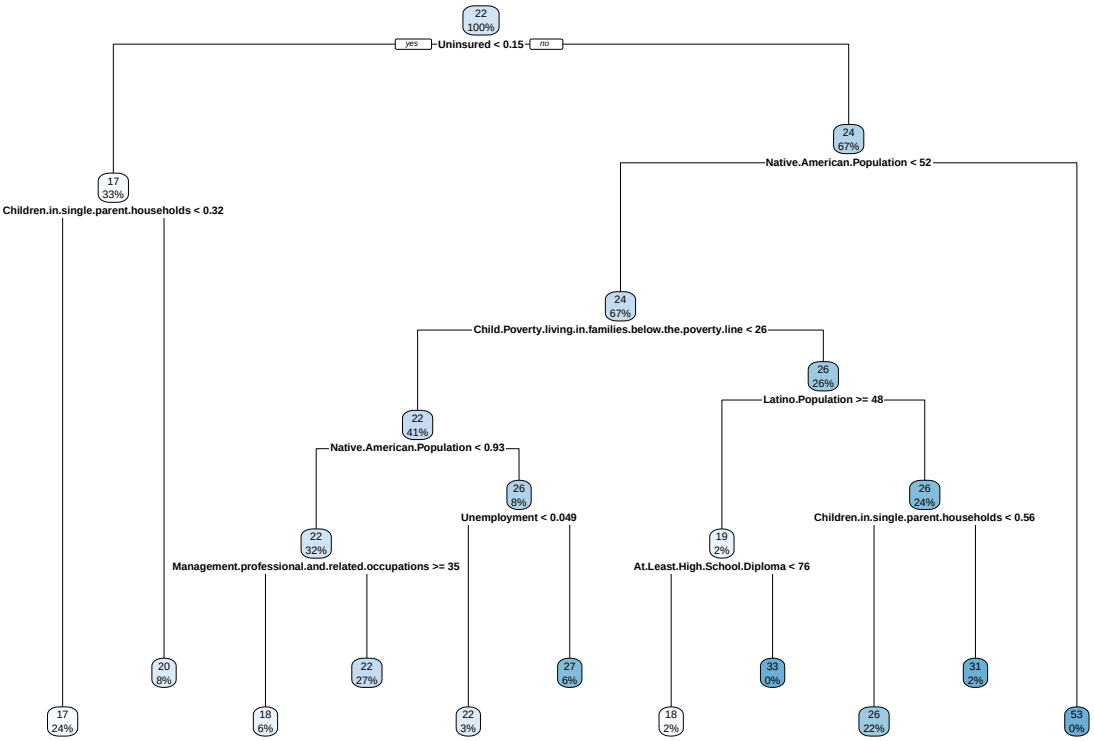

```
## [1] 14.23563
```

k-Nearest Neighbors mean square error

```
## [1] 13.94861
```

SELF: Self-harm and interpersonal violence

Regression tree and mean square error



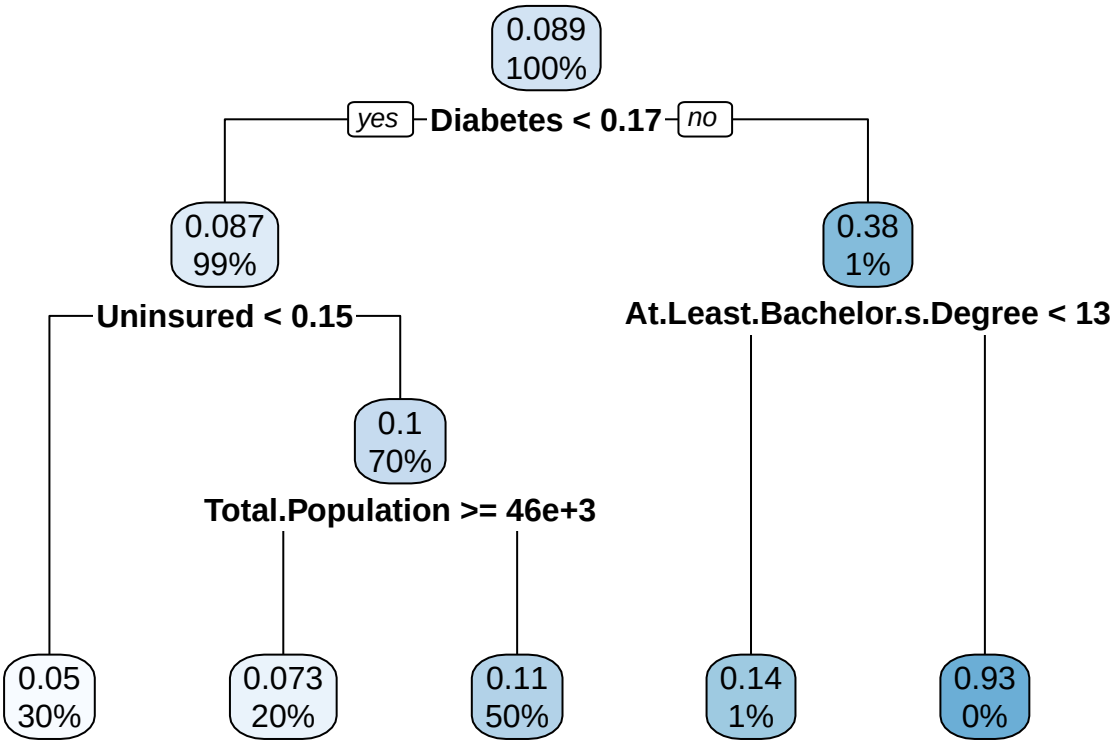

```
## [1] 25.32089
```

k-Nearest Neighbors mean square error

```
## [1] 29.88253
```

WAR: Forces of nature, war, and legal intervention

Regression tree and mean square error




```
## [1] 0.01477997
```

k-Nearest Neighbors mean square error

```
## [1] 0.01402043
```

Shiny Visualization

Our Shiny application can be found at https://teammortality.shinyapps.io/CountyMortality/. This portion is what we learned beyond what we covered in class.

The first important step to building our shiny application was to clean the data in a way that would be "function friendly." While researching Shiny and finding ways to build interactive maps of US county data, we discovered a package called choroplethr. From this package we utilize a function called county_choropleth() which builds a choropleth map of the counties in the United States. However, count-choropleth() takes a datafram with specific column names: region and value. Region must be assigned to the standard fips codes for the United States counties without any leading zeros. Value is then the information that will be projected onto the map. By setting number of colors to 1 in this instance we are able to create a continuous scale for our variable of interest - mortality rates. However, this particular function that county-choropleth() takes means that we need to manipulate our data set in several ways. This required us to rename our fips column to region and our mortality column to value.

Most importantly we needed to filter the data so that the values used when rendering a specific instance of our map could be selected by the user. To do this we created two drop down menus in the applet that would prompt filtering our master data set. Finally, we renamed the columns in our filtered data set to have a "region" and "value" column before using this new data set to build our map.

The master data set was bound using the following code:

Our ui.R page:

```
library(shiny)
library(maps)
library(ggplot2)
library(dplyr)
library(choroplethr)
library(googlesheets)
library(choroplethrMaps)


results <- read.csv("results.csv")

# Define UI for application that draws a histogram
shinyUI(fluidPage(

  titlePanel("County Level Mortality"),

  sidebarLayout(
    sidebarPanel (

    helpText("See how accurately our predictive models are able to predict
             the mortality rate for a given cause of death."),

    selectInput("var",

                label = "Choose Cause of Death",

                choices = list("HIV", "Maternal Disorders", "Common Infectious Diseases",
                               "Neglected Tropical Diseases", "Neonatal Disorders", "Nutritional Defici
                               "Other Communicable, Maternal, Neonatal and Nutritional Diseases",
                               "Neoplasms", "Chronic Respiratory Diseases", "Cirrhosis and Chronic Live
                               "Digestive Diseases", "Neurological Disorders", "Mental and Substance Us
                               "Diabetes, Urogenital, Blood and Endocrine Diseases", "Musculoskeletal
```

```
                                "Other Non-communicable Diseases", "Transport Injuries", "Unintentional
                                "Self-harm and Interpersonal Violence", "Forces of Nature, War, and Legal

                selected = "HIV"),


    selectInput("model",

                label = "Choose Model",

                choices = list("Actual Mortality Rates per 10,000", "Regression Tree Estimate Error", "kN

                selected = "Actual Mortality Rates")
    ),


  mainPanel(plotOutput("map"), width = 12,
            h6("*For model estimates, our legend refers to error rates. i.e. a '1.5' means our model ove
    )
  )
)
```

And our server.R page:

```
library(shiny)
library(maps)
library(ggplot2)
library(dplyr)
library(choroplethr)
library(googlesheets)
library(choroplethrMaps)

#county <- map_data("county")

results <- read.csv("results.csv")

shinyServer(function(input, output) {

  output$map <- renderPlot({

    cause.of.death <- switch(input$var,
                   "HIV" = "HIV",
                   "Maternal Disorders" = "MAT",
                   "Common Infectious Diseases" = "INF",
                   "Neglected Tropical Diseases" = "TROP",
                   "Neonatal Disorders" = "NEON",
                   "Nutritional Deficiencies" = "NUT",
                   "Other Communicable, Maternal, Neonatal and Nutritional Diseases" = "OTHC",
                   "Neoplasms" = "NEOP",
                   "Chronic Respiratory Diseases" = "CHRON",
                   "Cirrhosis and Chronic Liver Diseases" = "CIRR",
                   "Digestive Diseases" = "DIG",
                   "Neurological Disorders" = "NEUR",
                   "Mental and Substance Use Disorders" = "MENSUB",
```

```
                    "Diabetes, Urogenital, Blood and Endocrine Diseases" = "DIAB",
                    "Musculoskeletal  Disorders" = "MUSC",
                    "Other Non-communicable Diseases" = "OTHN",
                    "Transport Injuries" = "TRAN",
                    "Unintentional Injuries" = "UNIN",
                    "Self-harm and Interpersonal Violence" = "SELF",
                    "Forces of Nature, War, and Legal Intervention" = "WAR")

   model.choice <- switch(input$model,
                    "Actual Mortality Rates per 10,000" = "real",
                    "Regression Tree Estimate Error" = "tree",
                    "kNN Estimate Error" = "knn")


    results.to.use <- results %>% filter(model == model.choice) %>% filter(cause == cause.of.death) %>%
                              mutate(region = fips) %>% mutate(value = mortrate)

    county_choropleth(results.to.use, num_colors = 1)

  })

})
```

Limitations

The limitations of our project are five fold.

First the data sets we used contain several holes. For example, every county in Alaska as well as Oglala Lakota in South Dakota were missing mortality data. For this reason they show up as black spaces in our map.

Our second problem is inherent in mortality measures. Because mortality is measured on a scale of deaths per 10,000 people, our data easily drastically overestimates the mortality of counties with originally very small mortality rates. For example, if a county has close to 0 or 1 death per 10,000 and our estimate for that county is 3 deaths, that means our estimate is 300% off, despite the difference only being a person in 5,000. Next, none of our findings are causal. We show correlations and connections between similar counties and are able to highlight counties that differ vastly from the other counties most similar to them but we don't offer any information about why these counties are so different from their neighbors.

Penultimately, the R squared for our models are highly inconsistent meaning the reliability of our estimates and the information gained from them is variable. Our R squared ranges from as low as .08 to .67.

Finally, our model offers limited policy implications. Many of the factors that are correlated with mortality rates, such as racial composition and electoral results, cannot be affected by government action, while others, such as the sector composition of employment and educational attainment, can only be affected indirectly and a small amount. This limits the potential actual usefulness of this study to policymakers.

Future Extensions

There are a few directions in which we would extend or work given more data and time.

The first is the inclusion of data for more factors about the counties. In particular, we think it would be really interesting to look at the impact on mortality rates of the medical system in a given county - basic factors such as the number of doctors and hospital beds per capita, but also the quality of the care provided.

Next, we would include more types of statistical models if possible. For instance, we might build random forests or bagged trees instead of only k-nearest neighbors and regression tree models.

Finally, with more time we would use the dataset we already have to model and visualize trends in mortality rates over time. The mortality rates dataset we have has rates for each cause for each county every 5 years between 1980 and 2014. We did not have time to build models and visualizations for all of those years, focusing only on 2014 instead, but that would likely be a very interesting and insightful extension.