Community Health Status Indicators 2009

Ankita Tapadia & Prateek Chitpur

Applied Statistics & Visualization for Analytics

FALL-2019

George Mason University
STAT515-003-P01

Project Description

In order to analyze, visualize and build models on a health domain data, dataset has been picked from catlog.data.gov. The data belongs to CDC (Centers for Disease Control and Prevention) [1] who developed this data under the CHSI (Community Health Status Indicators) Project which provides Health Indicator on County level that tells the story about each county's Health. Here is the Link to download dataset.

Data Variables

The dataset file has various csv files, out of which we have picked two csv files: summarymeasuresofhealth and demographics. There are around 200 variables out of which the below mentioned variables are taken into main consideration. County-based data with variables as **FIPS (Federal Information Processing Standards) State and County code** which when combined forms a unique key. It has C**ounty names, State abbreviation**. As the data is based at county level, hence the **Unit of Analysis** could be either **County or State.**

It has the following main number variables of each county as:

ALE (Average Life Expectancy): Average number of years that a baby born in a particular year is expected to live. (Ratio)
All_Death: This is the period rate of death (any cause). (Ratio)
Unhealhty_Days: The average number of unhealthy days (mental or physical) in past 30 days. (Ratio)
Population Size: Mid-year estimates of resident population for 2008. (Interval)
Population Density: These field is calculated by using the formula: Population size divided by 2000 Land Area (square miles). (Interval)
Population Rates: The percentage of individuals living in each county is also provided by Race/Ethnicity (White, Black, Hispanic, Asian, Native American) and by age group of (under 19, 19 to 64, 64 to 85, above 85). (Ratio)
Health Status: For adults aged 18 or above, Status was set as 'poor' or 'fair' health which then converted to rate per each county. (Ratio)

Associated Research Questions

- Can a model that contains a significant number of predictors be developed, that reliably predicts the Health Status?
- How accurate and reliable are these models?
- Other questions have been mentioned and answered with the Visualizations in support in the further sections.

Data Tidying

The data from both the csv files have been merged and stored into one csv file. Dataset looks all clean and tidy, but it has also been specified that -1111 or -2222 values are already set for no data available or no report respectively. Such rows have not been imputed with any mean or median and just removed from the further observation as we did not want to impute a random value in already derived values of observations. For analysis 2431 observations are considered.

Exploratory Data Analysis

Tableau software provides a powerful data visualization platform in simplifying large raw data into very easy understandable format. So, this software is used for data exploration in this project.

- Is there a variation in population percentage of age group across the States?

  Figure 1 shows a graph which is plotted between population of different age group and CHSI State abbreviations. On comparison, it is observed that population of age between 19-64 is found to be the highest in the population rate among majority of the states and population of age between 85 and over is with the lowest population rate in all the states.

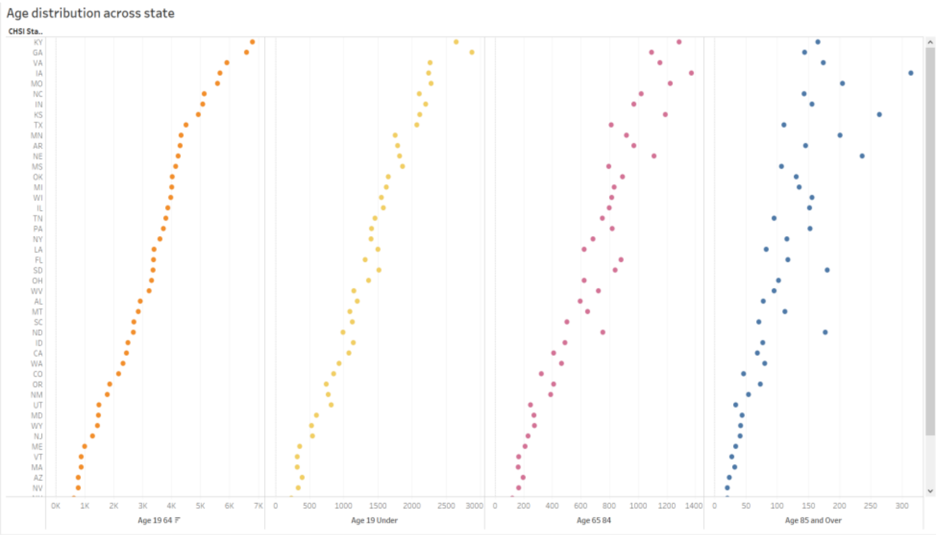


Figure 1: Graph shows population percentage of age group across states

- How are the States affected by the total death rate?

  A heat map is plotted between sum of all death of states and CHSI State abbreviations, as shown in *Figure 2*. This visual representation depicts that sum of all deaths is highest in Hawaii (HI), Kentucky (KY) and Georgia (GA) when compared to other states.

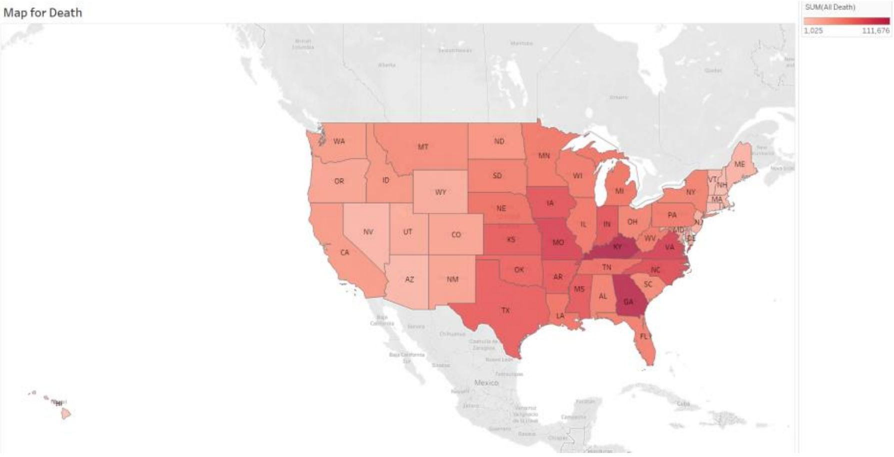


Figure 2: Heat map shows sum of all death across States

- Can population percentage by age group be categorized based on the Death rate?

Figure 3 illustrates a graph, which is plotted between sum of all death of states and total population percentage of different age group. It is interesting to note that sum of all death of different age group is found to be the same, with huge variation in the population range. We could see that for the least population percentage of age 85 and above also the death rate remains high and almost like the age 19-64 which has highest population rate.

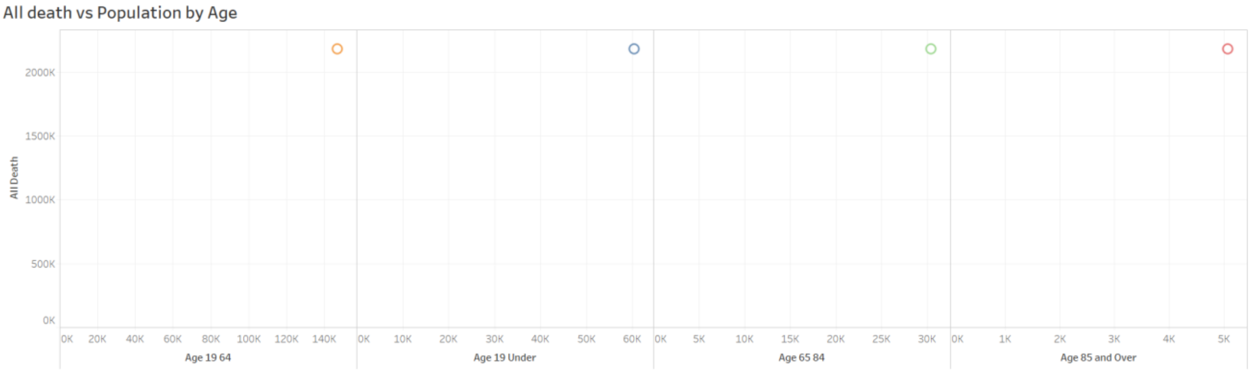


Figure 3: Graph showing all death against different population percentage by age group

- Does origin of person (White/black/Hispanic/Asian) varies across the U.S.A?

A bar graph is plotted between population rate of different race distributions and CHSI state abbreviations, which is shown in *Figure 4*. Population rate for the Asian is found to be highest in California (CA), for the Black in Georgia (GA), for the Hispanic in Texas (TX), for the Native American in South Dakota (SD) and for the White in Kentucky (KY).

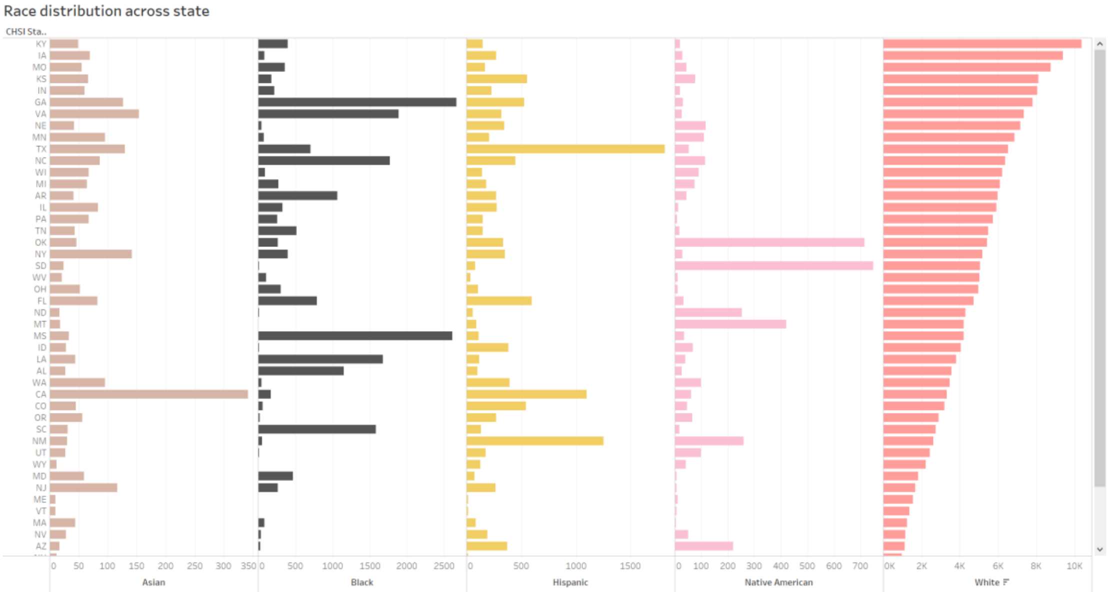


Figure 4: Bar graph showing population rate of different race distribution across states

- Is overall population playing a role in the all death causes rate?

A scatterplot is plotted between sum of all death and sum of population, as shown in *Figure 5*. It is clearly visible that as population increases, death also increases across all the States.

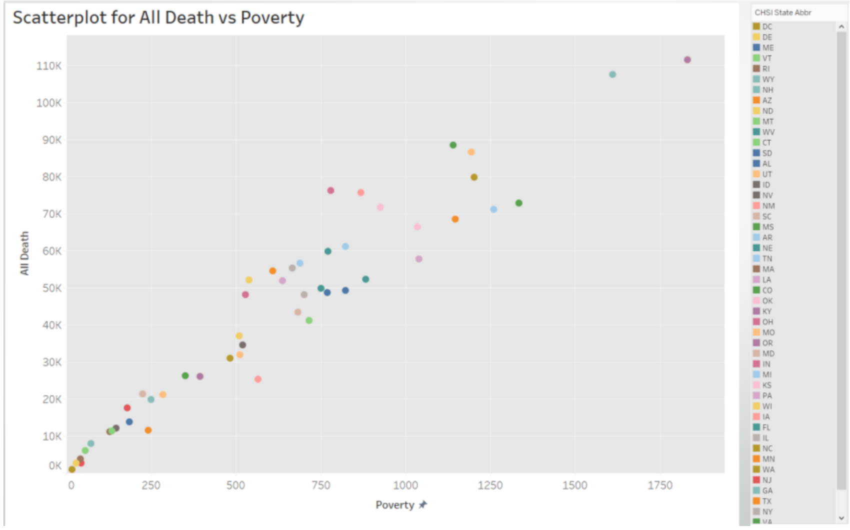
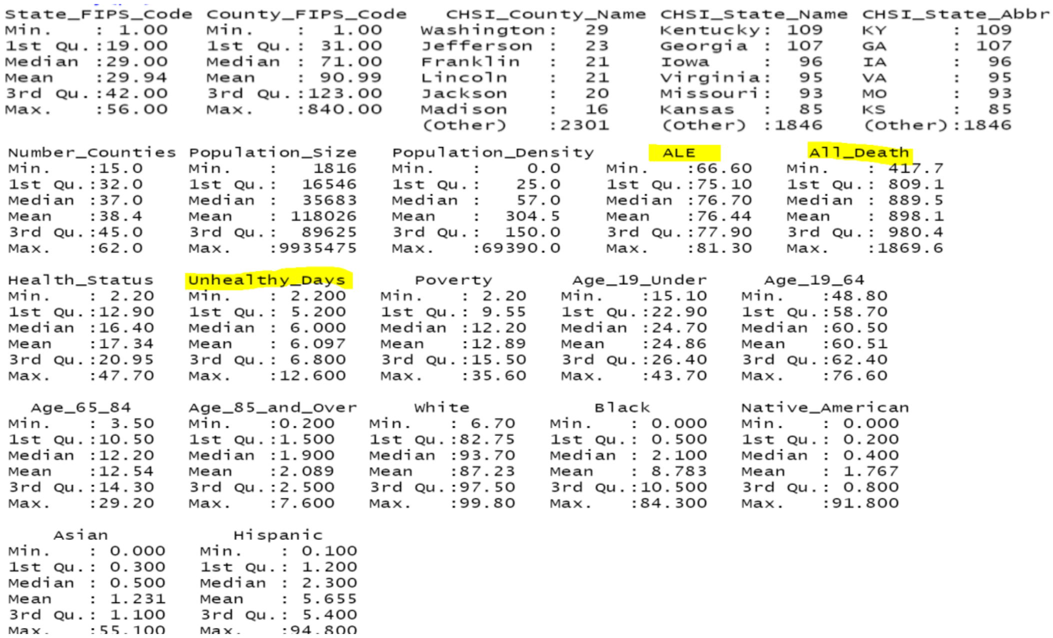


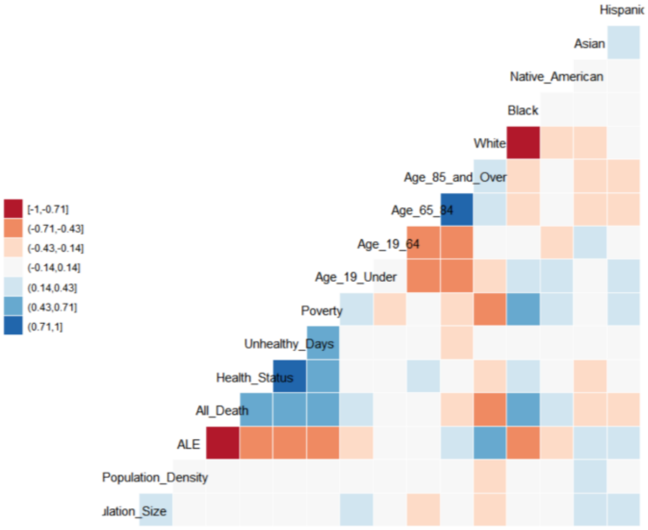
Figure 5: Scatterplot showing all death against population rate

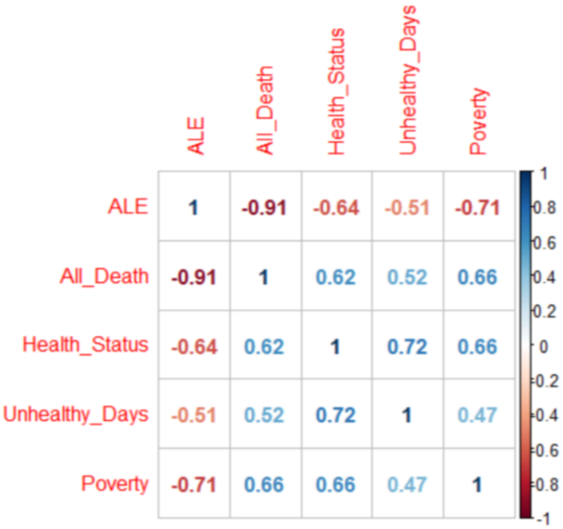
Data Analysis in R

After cleaning and visualizing data in Tableau, first a Tukey's 5-point summary function is run on the dataset, which gives us the minimum, 1st and 3rd Quartile, mean, median and maximum of each individual variable as follows:


```
State_FIPS_Code  County_FIPS_Code     CHSI_County_Name  CHSI_State_Name  CHSI_State_Abbr
Min.   : 1.00    Min.    :  1.00    Washington:   29    Kentucky: 109    KY      : 109
1st Qu.:19.00    1st Qu.: 31.00     Jefferson :   23    Georgia : 107    GA      : 107
Median :29.00    Median : 71.00     Franklin  :   21    Iowa    :  96    IA      :  96
Mean   :29.94    Mean   : 90.99     Lincoln   :   21    Virginia:  95    VA      :  95
3rd Qu.:42.00    3rd Qu.:123.00     Jackson   :   20    Missouri:  93    MO      :  93
Max.   :56.00    Max.   :840.00     Madison   :   16    Kansas  :  85    KS      :  85
                                    (Other)   :2301     (Other) :1846    (Other):1846

Number_Counties  Population_Size   Population_Density      ALE           All_Death
Min.   :15.0     Min.   :   1816   Min.   :    0.0    Min.   :66.60   Min.   : 417.7
1st Qu.:32.0     1st Qu.:  16546   1st Qu.:   25.0    1st Qu.:75.10   1st Qu.: 809.1
Median :37.0     Median :  35683   Median :   57.0    Median :76.70   Median : 889.5
Mean   :38.4     Mean   : 118026   Mean   :  304.5    Mean   :76.44   Mean   : 898.1
3rd Qu.:45.0     3rd Qu.:  89625   3rd Qu.:  150.0    3rd Qu.:77.90   3rd Qu.: 980.4
Max.   :62.0     Max.   :9935475   Max.   :69390.0    Max.   :81.30   Max.   :1869.6

Health_Status    Unhealthy_Days      Poverty         Age_19_Under     Age_19_64
Min.   : 2.20    Min.   : 2.200    Min.   : 2.20   Min.   :15.10   Min.   :48.80
1st Qu.:12.90    1st Qu.: 5.200    1st Qu.: 9.55   1st Qu.:22.90   1st Qu.:58.70
Median :16.40    Median : 6.000    Median :12.20   Median :24.70   Median :60.50
Mean   :17.34    Mean   : 6.097    Mean   :12.89   Mean   :24.86   Mean   :60.51
3rd Qu.:20.95    3rd Qu.: 6.800    3rd Qu.:15.50   3rd Qu.:26.40   3rd Qu.:62.40
Max.   :47.70    Max.   :12.600    Max.   :35.60   Max.   :43.70   Max.   :76.60

  Age_65_84      Age_85_and_Over       White            Black         Native_American
Min.   : 3.50    Min.   :0.200    Min.   : 6.70   Min.   : 0.000   Min.   : 0.000
1st Qu.:10.50    1st Qu.:1.500    1st Qu.:82.75   1st Qu.: 0.500   1st Qu.: 0.200
Median :12.20    Median :1.900    Median :93.70   Median : 2.100   Median : 0.400
Mean   :12.54    Mean   :2.089    Mean   :87.23   Mean   : 8.783   Mean   : 1.767
3rd Qu.:14.30    3rd Qu.:2.500    3rd Qu.:97.50   3rd Qu.:10.500   3rd Qu.: 0.800
Max.   :29.20    Max.   :7.600    Max.   :99.80   Max.   :84.300   Max.   :91.800

    Asian            Hispanic
Min.   : 0.000    Min.   : 0.100
1st Qu.: 0.300    1st Qu.: 1.200
Median : 0.500    Median : 2.300
Mean   : 1.231    Mean   : 5.655
3rd Qu.: 1.100    3rd Qu.: 5.400
Max.   :55.100    Max.   :94.800
```

Summary function also gives us good information about each variable. Next, Correlation plot is been plotted for each numerical variable as in the below first figure. This figure tells us that there are many variables like population size and density which are not strongly co-related with other variables. Hence a 2nd correlation plot has been plotted with the main or strongly co-related variables.

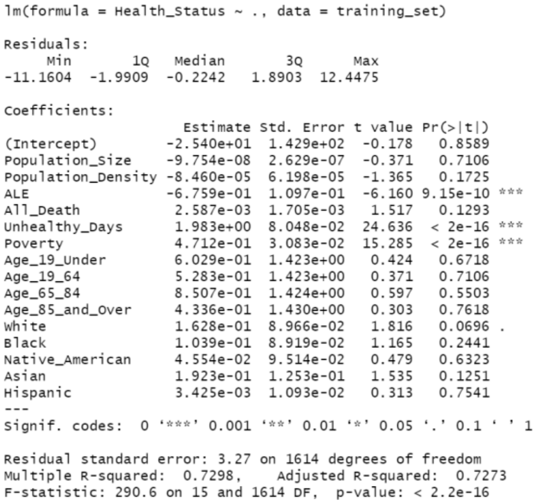




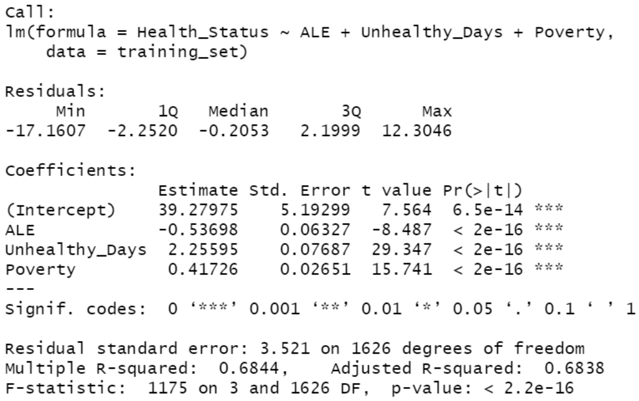
Model 1 - Linear Regression

To start building a model, using split ratio as 2/3 the data is split into training and test data set that divides the train data as 80% and test data as 20%. A multiple linear model is built using Health status as the response or target variable. While all other variables in the dataset is considered as the predictors. Once the model is built, using the summary function we can see what variables or predictors are significant.

Here we could see that only ALE, Unhealhty_Days and Poverty has '***' for being significant. Now using these 3 values a new model is built. The p-value is very less, which is good.

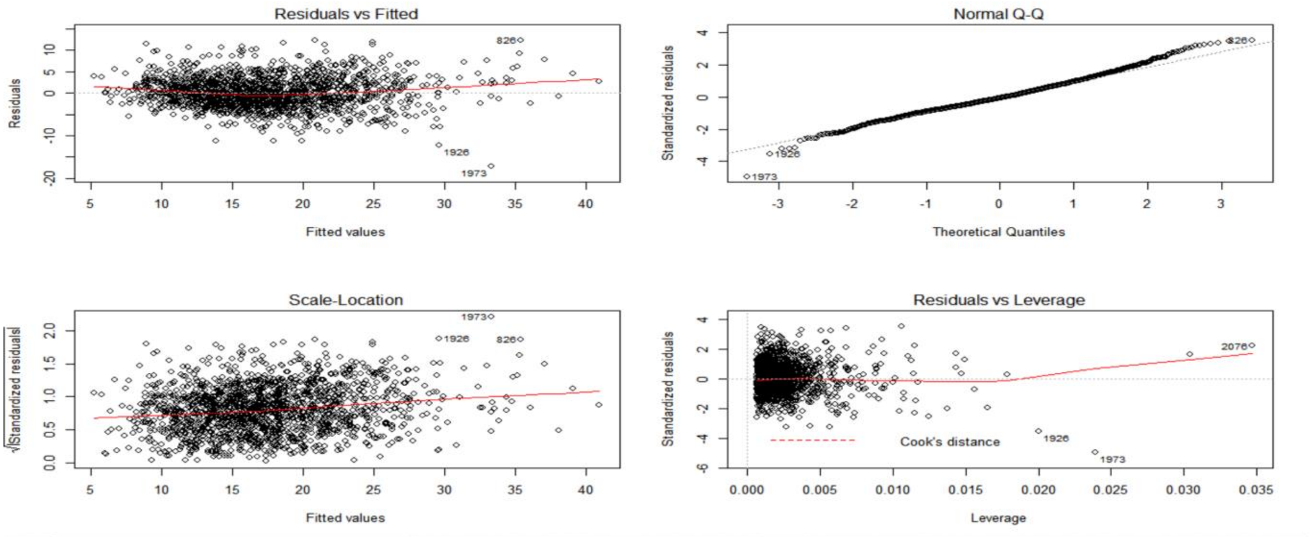


```
lm(formula = Health_Status ~ ., data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1604  -1.9909  -0.2242   1.8903  12.4475 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -2.540e+01  1.429e+02  -0.178   0.8589    
Population_Size    -9.754e-08  2.629e-07  -0.371   0.7106    
Population_Density -8.460e-05  6.198e-05  -1.365   0.1725    
ALE                -6.759e-01  1.097e-01  -6.160 9.15e-10 ***
All_Death           2.587e-03  1.705e-03   1.517   0.1293    
Unhealthy_Days      1.983e+00  8.048e-02  24.636  < 2e-16 ***
Poverty             4.712e-01  3.083e-02  15.285  < 2e-16 ***
Age_19_Under        6.029e-01  1.423e+00   0.424   0.6718    
Age_19_64           5.283e-01  1.423e+00   0.371   0.7106    
Age_65_84           8.507e-01  1.424e+00   0.597   0.5503    
Age_85_and_Over     4.336e-01  1.430e+00   0.303   0.7618    
White               1.628e-01  8.966e-02   1.816   0.0696 .  
Black               1.039e-01  8.919e-02   1.165   0.2441    
Native_American     4.554e-02  9.514e-02   0.479   0.6323    
Asian               1.923e-01  1.253e-01   1.535   0.1251    
Hispanic            3.425e-03  1.093e-02   0.313   0.7541    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.27 on 1614 degrees of freedom
Multiple R-squared:  0.7298,	Adjusted R-squared:  0.7273 
F-statistic: 290.6 on 15 and 1614 DF,  p-value: < 2.2e-16
```

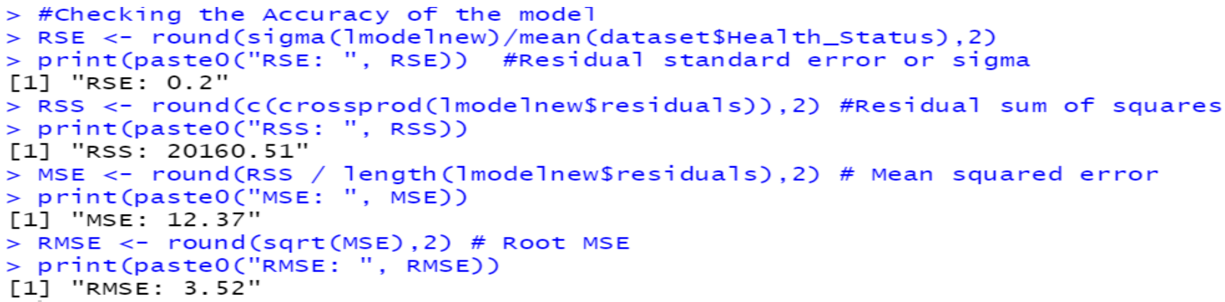


```
Call:
lm(formula = Health_Status ~ ALE + Unhealthy_Days + Poverty, 
    data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.1607  -2.2520  -0.2053   2.1999  12.3046 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    39.27975    5.19299   7.564  6.5e-14 ***
ALE            -0.53698    0.06327  -8.487  < 2e-16 ***
Unhealthy_Days  2.25595    0.07687  29.347  < 2e-16 ***
Poverty         0.41726    0.02651  15.741  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.521 on 1626 degrees of freedom
Multiple R-squared:  0.6844,	Adjusted R-squared:  0.6838 
F-statistic:  1175 on 3 and 1626 DF,  p-value: < 2.2e-16
```

Now, that the model has significant variables as predictors, let’s check further summary stats and linearity assumptions.



From the residuals and fitted graph plots we could see that it doesn’t satisfy all the linearity assumptions [2]. In the Normal Q-Q plot, the points are little deviated from the upper most line. In the cook’s distance leverage graph, there are many outliers, but not all outliers should be removed. In the Skewed plot, the data points are spread across. In the 1st residual vs fitted value plot; the red line is not straight. Hence, we could see that the linear model doesn’t fit well. From the numeric values also, we could figure it out.



```
> #Checking the Accuracy of the model
> RSE <- round(sigma(lmodelnew)/mean(dataset$Health_Status),2)
> print(paste0("RSE: ", RSE))  #Residual standard error or sigma
[1] "RSE: 0.2"
> RSS <- round(c(crossprod(lmodelnew$residuals)),2) #Residual sum of squares
> print(paste0("RSS: ", RSS))
[1] "RSS: 20160.51"
> MSE <- round(RSS / length(lmodelnew$residuals),2) # Mean squared error
> print(paste0("MSE: ", MSE))
[1] "MSE: 12.37"
> RMSE <- round(sqrt(MSE),2) # Root MSE
> print(paste0("RMSE: ", RMSE))
[1] "RMSE: 3.52"
```

Here above the mean square error is 12.37 for the linear model. We could see that model is not reliable and accurate. When linear model was built using ALE (Average life Expectancy) as the response variable and others as predictor, the values came out to be more reliable and accurate. This has been added in the Appendix section and is present in R code.

Model 2 - Random Forest

Random forest is an ensemble learning method for regression, classification which functions by constructing many decision trees. It gives better prediction when compared to other classifiers. The number of attributes selected by random forest in building trees does not affect the overall prediction value. It handles missing values. Also, random forest avoids overfitting. Hence, random forest is selected in building model for prediction. The target variable in our model is Health_Status.

Here are the packages and libraries used in building this model, randomForest [4], caret [3], 'e1071' and randomForest, caret respectively.

The dataset file ‘projectdata.csv’ file is read using following code.

```
rfdata <- read.csv("projectdata.csv", sep = ",")
```

Health_Status attribute is converted into a factor variable in building model. So, the random forest model here is a classification model.

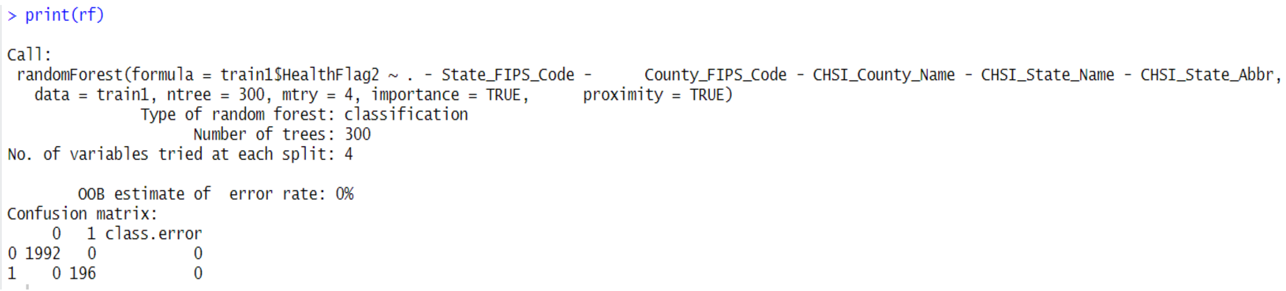
```
rfdata$HealthFlag <- with(rfdata, rfdata$Health_Status >= 25)
rfdata$HealthFlag2 <- as.integer(rfdata$HealthFlag)
rfdata$HealthFlag2 <- as.factor(rfdata$HealthFlag2)
```

The data frame ‘rfdata’ is divided into train and test data in the ratio 70:30 respectively.

Below is the random forest model code built and trained with train1 data, considering ‘HealthFlag2’ as target variable against all attributes of dataset as predictors, except categorical variables ‘State_FIPS_Code’, ‘County_FIPS_Code’, ‘CHSI_County_Name’, ‘CHSI_State_Name’, ‘CHSI_State_Abbr’.

```
set.seed(222)
rf <- randomForest(train1$HealthFlag2 ~.-State_FIPS_Code -County_FIPS_Code -CHSI_County_Name -CHSI_State_Name -CHSI_State_Abbr, data = train1, ntree = 300, mtry = 4, importance = TRUE, proximity = TRUE)
```

The print() gave the following result.



```
> print(rf)

Call:
 randomForest(formula = train1$HealthFlag2 ~ . - State_FIPS_Code -      County_FIPS_Code - CHSI_County_Name - CHSI_State_Name - CHSI_State_Abbr,
   data = train1, ntree = 300, mtry = 4, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0%
Confusion matrix:
     0   1 class.error
0 1992   0           0
1    0 196           0
```

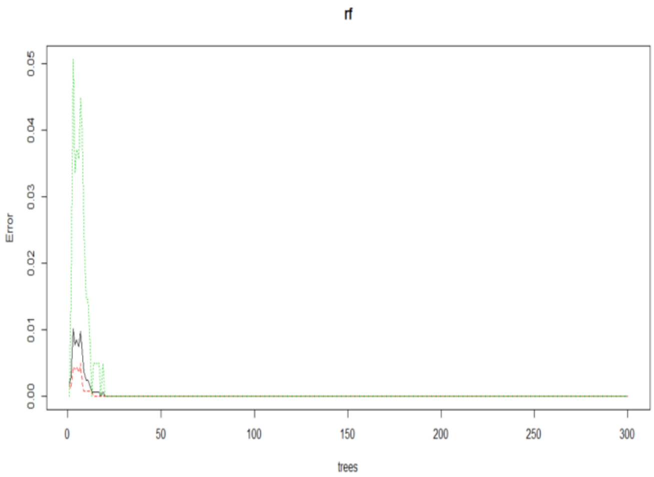
The Out of Bag (OOB) error rate shown 0%.

The random forest model (rf) is then tested with the test1 data, to check the prediction of the model.

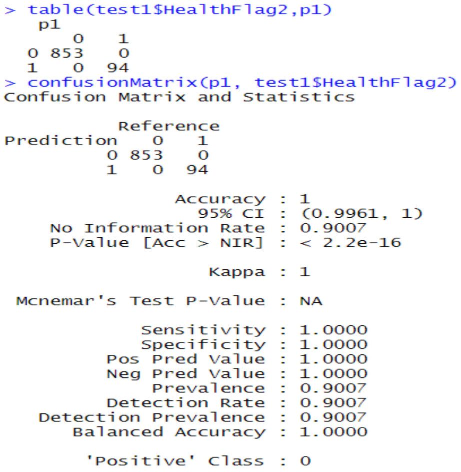
```
p1 <- predict(rf, newdata = test1)
```

The Out of Bag (OOB) error rate slumps down with the increase in number of trees in random forest model, which can be observed from the below plot.

plot(rf)



table() and confusionMatrix() results:

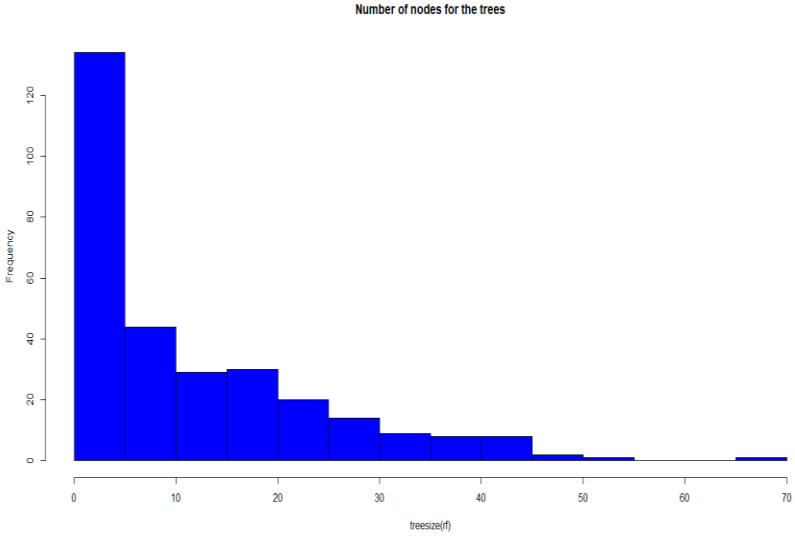


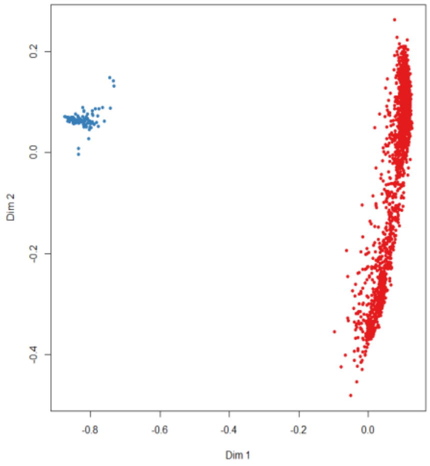
```
> table(test1$HealthFlag2,p1)
   p1
      0   1
  0 853   0
  1   0  94
> confusionMatrix(p1, test1$HealthFlag2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 853   0
         1   0  94

               Accuracy : 1
                 95% CI : (0.9961, 1)
    No Information Rate : 0.9007
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.9007
         Detection Rate : 0.9007
   Detection Prevalence : 0.9007
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

From these two functions we can see that the predicted and actual values gave us 853 and 94 true positives with no error, accuracy as 1 and p-value less than 2.2e-16, which is good.

Histogram shows the number of nodes for tress.

hist(treesize(rf), main = "Number of nodes for the trees", col = "blue")





Multidimensional plot shows an overview of similarities and differences of Health_Status.

MDSplot(rf, train1$HealthFlag2)

Challenges

During the conduct of our project, we encountered two major challenges. First, the dataset contained around 10 csv files, so selecting the meaningful dataset among those files required reading through all the documentation and understanding the terms pertaining to the CHSI dataset. Second, the dataset had no NA values, however there are some missing values and majority of the missing values (no data available or no report) are filled with either -1111 or -2222. Hence, much time was spent on data cleaning and tidying.

Conclusion

Thus, the CHSI (Community Health Status Indicator) as set by the data owner gives the major information required by any individual to find out county's health. This data set is developed by the CDC to help serve the community towards a better health, which is truly achieved by the reading and analysis of the data. Linear regression did not satisfy all the assumptions of linearity completely but still gives a good prediction rate. There exists a case of multicollinearity between the variables. Random forest gave better prediction when compared to other classification models. It worked well and gave an accuracy of 100% in predicting target variable Health Status.

Contribution

Data cleaning, Tableau visualization and report writing has been done by both with equal contributions. Model-1 (Linear regression) is done in R by Ankita Tapadia, while Model-2 (Random Forest) in R is done by Prateek Chitpur. While both have equally contributed in writing their models and reviewing each other's work. This gave us more clarification on each other's work too.

References

[1] Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer. (2019, September 2). Retrieved November 5, 2019, from https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer

[2] Analytics Vidhya Content Team. (2019, June 25). Regression Analysis with Assumptions, Plots & Solutions. Retrieved November 21, 2019, from https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

[3] Kuhn, M. (2008), "Building predictive models in R using the caret package, " Journal of Statistical Software, Retrieved from: http://www.jstatsoft.org/article/view/v028i05/v28i05.pdf

[4] Chaitanya Sagar, Prudhvi Potuganti and Saneesh Veetil. (2018), "How to implement Random Forests in R", R-bloggers, Retrieved from: https://www.r-bloggers.com/how-to-implement-random-forests-in-r/

Appendix A

While downloading the data from the source link, we also downloaded the PDF, which consist of all the data source, definition and notes. This file is important in order to understand each variable and the complete data.

PDF

CHSI-Data-Sources-Definitions-And-Notes-