Predicting Life Expectancy using Community Health Status Indicators (CHSI) Data

Group 8: Jie Yu, Shuwei Liu, and Noah Kreski

Introduction

Life expectancy is one of the most important measures of health. It serves as a meaningful health indicator that can be compared between countries or regions within countries (Mathers et al., 2001), and helps to answer the most fundamental question concerning health: how long can the typical person expect to live? Life expectancy may be the outcome of multiple factors, such as lifestyle choices, access to healthcare, socioeconomic status, community and social support (WHO, 2016).

According to the latest annual report from the National Center for Health Statistics (2018), the average American could expect to live 78.6 years, down from 78.7 years in 2016, and as a key indicator for the Sustainable Development Goal to “Ensure healthy lives and promote well-being for all at all ages” (WHO, 2016), prolonging lifespan is an urgent and necessary goal. Public health professionals need to ensure the efficient allocation of resources to combat this decline in life expectancy. Under such circumstances, prediction of life expectancy is necessary, providing vital information for relevant health policy making.

In this project, we aim to build a model that 'best' predicts the average life expectancy per county in the United States, and examine the association between multiple factors, including those mentioned above, and average life expectancy.

Methods

Data Description and Preparation

We used the Community Health Status Indicators (CHSI) dataset which was released in 2012 by CDC. The thirty-three predictor variables we picked provide information at the county level related to residents’ sociodemographic characteristics, lifestyle choices, access to healthcare and other health-related risk factors. When we checked the correlations among predictor variables, we found that there was a group of eight variables measuring the total count of individuals for a county which were all highly correlated with county population size. We felt that proportions of key factors were more pertinent than pure counts, and so by dividing those variables by county population size, we successfully eliminated most of the correlations. Our final predictors include age, race, population size and density, death rate, percent with poor health status, number of monthly unhealthy days, percent not exercising, percent with few fruits or vegetables, percent obese, percent with high blood pressure, percent smoking, percent with diabetes, physician rate, dentist rate, presence of a community health center, health professional shortage (HPSA) designation, percent poverty, percent unemployed, percent uninsured, percent medicare based on elderly status and disability, percent with no high school diploma, percent severely disabled, percent with major depression, and percent with recent drug use. Twelve predictor variables have missing data, and we decided to perform the K-nearest neighbors (KNN) imputation on them in the subsequent model building part.

Our response variable of interest, average life expectancy, is also provided by this data set at the county level. We excluded two observations with missing values for the response, therefore we had a total of 3139 county-level observations.

Exploratory Data Analysis / Unsupervised Analysis

The descriptive statistics for all continuous variables is shown in Table 1. There were two categorical variables, HPSA designation which was seen in 75.95% of the sample, and presence of a community health center, 55.11% of the sample.

According to the correlation plot (Figure 1), some of the variables in our data were highly correlated with one another. Percent with age 85+, percent with age 65-84, and percent medicare based on elderly status show strong positive correlations, which makes sense because they are all related to the elderly. Percent uninsured and percent poverty are positively correlated as individuals below poverty are at high risk of being uninsured. Number of monthly unhealthy days, percent with poor health status, percent medicare based on disability, percent with no high school diploma, and percent not exercising are positively correlated with one another, which may provide redundant information. Besides, it is easy to understand that percent white and percent black show a strong negative correlation.

Additionally, hierarchically clustering using complete linkage and Euclidean distance (Figure 2) combined with a heatmap (Figure 3) were utilized to better understand the health profiles of regions in relation to one another. In order to make the visualizations readable, here all our county-level data are aggregated to the county level and take an average since counties within a state often share similarities.

When clustering the states into five clusters, a number of valuable results are found. For one, the first split separates Washington DC from all other states, likely due to a substantially higher pattern of healthcare access. Washington DC overwhelmingly has the highest primary care physician rate and dentist rate according to the heatmap, and one of the lowest rate of being uninsured. This area may not need the same interventions targeting healthcare infrastructure that other regions need.

The second cluster, including states such as Mississippi, North Carolina, Louisiana, and Oklahoma, features states that exclusively resides in the South, primarily the South-Eastern United States. This seems to reflect patterns of elevated negative health behaviors and outcomes. This region shares increased poverty, obesity, lower education, limited exercise, and increased smoking and diabetes, all of which contributes to the highest death rates and lower life expectancy. This cluster should be the priority for any interventions, as it seems to have the most urgent need for assistance.

The third cluster is just the state Hawaii, which features the lowest death rate and the highest life expectancy of any state. The characteristics seem reversed compared to the second cluster, as Hawaii exhibits lower smoking and obesity, and higher exercise and education. Even without the substantial access to care seen in Washington DC, this area is living longer, healthier lives. Further inquiry should examine what practices contribute to Hawaii's success.

The fourth cluster includes Florida, Iowa, Nebraska, North Dakota, and South Dakota, which features a distinctly older population with higher proportions of citizens age 65 and older, plus more elderly people on medicare.

The remaining cluster features 33 states and further divisions, most of which follow some geographic logic, such as the six states of New England being grouped together. The overall results here are that there are distinct clusters of states, many of which are geographically similar, with unique patterns of health behaviors and resources, and this information can inform the development and implementation of strategies to improve life expectancy.

Models

Multiple models were examined to choose the best one, including ridge regression, LASSO, principal components regression (PCR), partial least squares regression (PLS), generalized additive model (GAM), multivariate adaptive regression splines (MARS), K-nearest neighbors (KNN), and tree-based methods including regression tree, bagging, random forests and boosting. An ordinary least squares regression (OLS) was also prepared to compare with other models. Ultimately, random forest was isolated as the most 'optimal' model to fit the data.

A design matrix of predictors was prepared, and data was split into training and test datasets with 75% chosen as the training rows. Scaled, centered and KNN-imputed data was used to build models during cross-validation period since there are a number of missing values in the original dataset.

Except for OLS, all the other models have tuning parameters in their model building. We utilized 10-fold cross validation to find the optimal tuning parameters which gave us the smallest cross-validation error, and refit the models on the training data using the optimal tuning parameters. Root mean squared error (RMSE) were calculated to compare the model performance by summary statistics (Table 2) and violin box plot (Plot 4). Models were also applied on the test dataset to make predictions for the response, and test mean squared error (MSE) were calculated for model validation.

Ridge: As one of the shrinkage methods, it can perform particularly well when there is a subset of true coefficients that are very small. We tested a sequence of tuning parameter λ from exp(-5) to exp(2), and the optimal one chosen by cross-validation was 0.385. In terms of mean RMSE (1.2115), ridge performed better than OLS, but not that well compared to others (Table 2).

LASSO: As another shrinkage method, it can penalize the coefficient estimates of unimportant predictors to zero and offer a clearer interpretation compared with ridge regression. A sequence of tuning parameter λ from exp(-4) to exp(1) was tested and the optimal one chosen was 0.053.The final model provided 15 non-zero coefficients, which might play important role in predicting average life expectancy. As expected, LASSO removed some highly correlated variables and kept the most essential ones, for example, keeping percent not exercising, and removing the variables highly correlated with it as mentioned above. Mean RMSE of this model (0.9033) is in the middle level among all the methods (Table 2).

PCR: This method was utilized to examine the predictors as linear combinations that reduce dimension and capture joint variation. Here, the response does not supervise this identification of these principal components. This method led to the identification of 25 principal components, based on cross-validated results that showed the smallest RMSE at this number, reduced from 33 predictors overall. In terms of mean RMSE (0.9018), PCR performed better than ridge and LASSO, but worse than MARS and tree-based ensemble methods (Table 2).

PLS: This method operates similar to PCR, but in a supervised way to ensure that new features are related to the response. Here, the optimal number of components was only 4, a severe reduction from the initial sample. While the mean RMSE from this model (1.0167) was larger than PCR, the PLS model offers helpful simplicity.

GAM: This model obtained its optimal smoothing parameter through generalized cross-validation. It extends the linear model to allow for certain non-linear relationships, was not only computationally intensive and prohibitively complex, but yielded a relatively large mean RMSE (1.1308) among all the models (Table 2). While it may serve some use in isolating non-linearity within the data, it does not serve as a useful model for these data in terms of practicality.

MARS: This method was used to create a piecewise linear model with an automatically chosen cut point. The best combination of tuning parameters were identified by a grid where the result showed that 11 degree of freedom allowing interactions in the model is appropriate (degree of features = 2, maximum number of terms including intercept = 11). The predictors chosen by MARS include percent with no high school diploma, death rate, percent native American, percent black, population size, percent poverty, percent obese, interaction between population size and death rate, interaction between native American and Hispanic, interaction between poverty and obesity, interaction between poverty and recent drug use, and interaction between native American and poverty. Mean RMSE of MARS (0.8783) is larger than tree-based ensemble methods but smaller than others (Table 2).

KNN: The KNN model was found to have an optimal tuning parameter of 8 with a tuning grid that first assessed a wide range of possible k values (from 1 to 300 by 5). Then, upon finding that RMSE increased from a k of 21 onward, the tuning grid was narrowed down (from 1 to 50 by 1). This yielded a mean RMSE of 0.9863, which is in the middle level among all the models (Table 2).

Regression Tree: Using a regression tree allows for the segmentation of the predictor space into simpler regions for regression. It requires a complexity parameter, which was tested from exp(-9) to exp(-5) and found through cross-validation to be approximately 0.001544, resulting in a tree with 35 terminal nodes. While its mean RMSE (1.0179) is much larger than those of ensemble methods (Table 2), it certainly lends itself to be highly interpretable. The first split occur using the death rate, which makes sense for a life expectancy outcome, and then goes into determinants like poverty and lack of exercise.

However, since the regression tree did not perform well, we tried some ensemble methods (bagging, random forest, and boosting) by aggregating a large number of decision trees in order to substantially improve the predictive performance.

Bagging: Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method by averaging highly-correlated variables within bootstrapped training datasets. We tried a wide range of possible node size of the tree (from 1 to 15). Cross-validation was applied to get the best tuning parameter which was 5. The mean RMSE equals to 0.859 which turns out to be smaller than any other method we tried above.

Random Forests: Random forests provides an improvement over bagged trees by decorrelating the trees. Although it is still based on bootstrapped training dataset, it reduces variance by using a random selection of feature at each splitting steps. Theoretically, a fresh selection of *m* predictions is taken at each split and m = p/3 would be chosen in this case. We tried a large range of m (from 1 to 33) and node size (from 1 to 15). Finally, the best tuning parameter appeared at m = 13 while minimum node size = 1. The mean RMSE is 0.830, which is the smallest one among all the methods (Table 2).

Boosting: Boosting works in a similar way as bagging, except that each tree is grown using information from previously grown trees. That is, we fit a decision tree using the current residuals from the model as the response, then we add new small trees into the fitted function and update the residuals. Using this strategy, we slowly improve the fitted function in areas where it does not perform well. There are four tuning parameters: the number of trees B, the shrinkage parameter λ (the learning rate), the interaction depth d (the maximum depth of each tree), and the minimum number of observations in each node. Boosting is sensitive to tuning and has potential to overfit, so we tuned it very carefully. We made several attempts, and in the last one, we set the the minimum number of observations in each node as 1, and tested a sequence of B from 4300 to 5000 by 100, tested λ from 0.003 to 0.005, and d from 8 to 16. The optimal value is 4900 for number of trees B, 0.05 for shrinkage parameter λ, and 15 for interaction depth d. The mean RMSE is 0.8329, which is quite small among all the methods, just slightly larger than random forests (Table 2). Based on this method, death rate is the most important predictor, followed by percent poverty and percent black.

Ultimately the most 'optimal' model is random forests, which produced the smallest cross-validation RMSE and was further validated with the smallest test MSE, 0.579. Boosting got the second smallest test MSE (0.586), which makes sense since it's performance was good and really close to that of random forests. KNN and PLS turned out to have the largest test MSE, which were 0.888 and 0.829 respectively. Considering their mean cross-validation RMSE were in a moderate level among all the methods, there might be an overfitting in those two models which we created on the training data.

In terms of what random forests, our best model, communicates about the outcome of average life expectancy, a few variables carry substantial importance. Naturally, the death rate is most important by far for life expectancy as a higher death rate means that people do not live as long. Following that, however, we have socioeconomic factors like poverty and lacking a high school diploma, medical and health-related experiences, such as the rate of disability-related medicare use and the percent with only fair or poor health status, and demographic factors such as the percent of a country's population who are Black.

Partial Dependence Plots (Figure 5) enable us to see the average marginal effect these factors have on the prediction of average life expectancy. For instance, with a death rate of about 700 per 100,000, average life expectancy stays above

78 years, but when the rate is 1,100 per 100,000, this plummets below 75 years. A 20 percentage point increase in the percent poverty drops the average life expectancy by about 0.7 years. As predictors, percent with no high school diploma, percent with lower health status, and percent black all exhibit a strong drop in life expectancy as they increase, but only on the scale of a few tenths of a year overall. Still, examining these factors is critical to the study of life expectancy. Random Forests provided a clear picture of key variables and the relationship that consistently explain average life expectancy at the county level.

We can also see the impact of these variables on a typical county, like Guthrie, Iowa. The feature effect plot here (Figure 6) articulates the role of each of the key factors on the prediction of average life expectancy generated by random forests. The relatively lower death rate in this county contributed heavily to its predicted average life expectancy of 78.69, followed by the relatively lower poverty rate. Limited percent with no high school diploma, racial demographics, limited lack of exercise, limited disability and high use of medicare among the elderly all carried positive, yet smaller, impacts on life expectancy. Seeing as Guthrie's real average life expectancy is 79.1, only 0.4 years off from the prediction, this model seems to accurately depict the experience of a healthy county. A less healthy county, such as Lamar, AL, also shows a clear depiction of what contributes to its lower predicted average life expectancy of 74.24, which is not far from its true value, 73.9. For the most part, the key variables are the same as a healthier county, just with different values (e.g. more poverty, fewer high school diplomas, higher death rate).

Conclusions

Random forests fit the data best in this study, and this model elaborates some meaningful findings. For example, poverty and education are absolutely essential components of the discussion on how to improve life expectancy, perhaps even more so than specific activities like smoking and exercise. We must acknowledge disparities where they exist, as seen with lower life expectancy predictions with increases in disability and percent Black. Besides, unsupervised learning allowed for the characterization of distinct clusters of states in ways that will expand our ability to thoughtfully approach the improvement of average life expectancy.

References

Mathers, C.D., Sadana, R., Salomon, J.A., Murray, C.J.L., Lopez, A.D (2001). Healthy life expectancy in 191 countries, 1999. Lancet 2001;357:1685-91.

National Center for Health Statistics (NCHS). Health, United States, 2017: With special feature on mortality. Hyattsville, MD, 2018. Retrieved April 5, 2019, from https://www.cdc.gov/nchs/data/hus/hus17.pdf

World Health Organization (WHO). World Health Statistics 2016: Monitoring the Health Goal - Indicators of Overall Progress, Retrieved March 11, from

https://www.who.int/gho/publications/world_health_statistics/2016/EN_WHS2016_Chapter3.pdf

Table 1: Descriptive Statistics for Continuous Variables (At the County Level)

Variable (Per County)	NAs	Mean	Std. Dev.	1st Quartile	Median	3rd Quartile
Average Life Expectancy	0	76.32	2	75	76.5	77.7
Population Density	1	250.07	1703.27	17	44	109.75
Population Size	0	94427.06	306520.4	11220	25270	64111
Percent with Age <19	0	24.81	3.28	22.7	24.6	26.4
Percent with Age 19-64	0	60.28	3.35	58.3	60.3	62.3
Percent with Age 65-84	0	12.79	3.33	10.7	12.5	14.7
Percent with Age 85+	0	2.12	0.95	1.5	1.9	2.6
Percent White	0	87.04	16.14	82.8	94.1	97.6
Percent Black	0	8.99	14.55	0.5	2.1	10.3
Percent Native American	0	1.95	7.62	0.2	0.4	0.9
Percent Asian	0	1.12	2.76	0.3	0.5	1
Percent Hispanic	0	7.02	12.47	1.1	2.3	6.3
Death Rate	3	905.64	131.21	814.25	898.6	989.8
Percent with Health Status	662	17.32	6.09	12.9	16.4	20.9
Number of Monthly Unhealthy Days	0	0.019	0.36	0.021	0.025	0.03
Percent not Exercising	933	26.51	6.7	21.9	26	30.8
Percent with Few Fruits or Vegetables	1235	78.92	5.16	75.5	79	82.4
Percent Obese	915	24.15	4.9	21.1	24.3	27.2
Percent with High Blood Pressure	1617	26.48	5.44	22.8	26.2	29.9
Percent Smoking	872	23.11	5.73	19.4	23	26.7
Percent with Diabetes	420	7.81	2.76	5.9	7.5	9.45
Physician Rate	0	57.6	44.78	30.55	50.6	74.7
Dentist Rate	1	32.19	21.5	18.7	30	43.3
Percent Poverty	1	13.35	4.88	9.8	12.6	16.2
Percent Unemployed	543	6.11	1.34	5.2	6	6.8
Percent Uninsured	0	0.14	0.36	0.11	0.13	0.17
Percent Medicare Based on Disability	0	0.017	0.79	0.017	0.023	0.031
Percent Medicare Based on Elderly Status	0	0.098	0.8	0.11	0.14	0.17
Percent with Major Depression	0	0.061	0.0066	0.056	0.059	0.065
Percent with No High School Diploma	0	0.15	0.058	0.11	0.14	0.19
Percent with Recent Drug Use	0	0.052	0.011	0.045	0.051	0.057
Percent Severely Disabled	0	0.025	0.062	0.02	0.027	0.034

Table 2: Model Comparison by Cross-validation RMSE

Model	Mean RMSE	Variance RMSE	1st Quantile	Median RMSE	3rd Quantile
Least Squares	1.3060	1.7205	0.8451	0.8839	0.9840
Ridge	1.2115	0.9278	0.8545	0.9064	0.9938
GAM	1.1308	0.1676	0.9819	1.0259	1.0550
Regression Tree	1.0179	0.0166	0.9609	1.0091	1.0925
PLS	1.0167	0.1299	0.8595	0.8855	1.0013
KNN	0.9863	0.0088	0.9094	1.0191	1.0444
LASSO	0.9033	0.0099	0.8668	0.8993	0.9601
PCR	0.9018	0.0103	0.8485	0.8789	0.9788
MARS	0.8783	0.0115	0.8053	0.8676	0.9717
Bagging	0.8594	0.0177	0.7742	0.8874	0.9274
Boosting	0.8329	0.0153	0.7300	0.8510	0.9340
Random Forest	0.8309	0.0142	0.7490	0.8502	0.8946

Figure 1: Correlation Plot

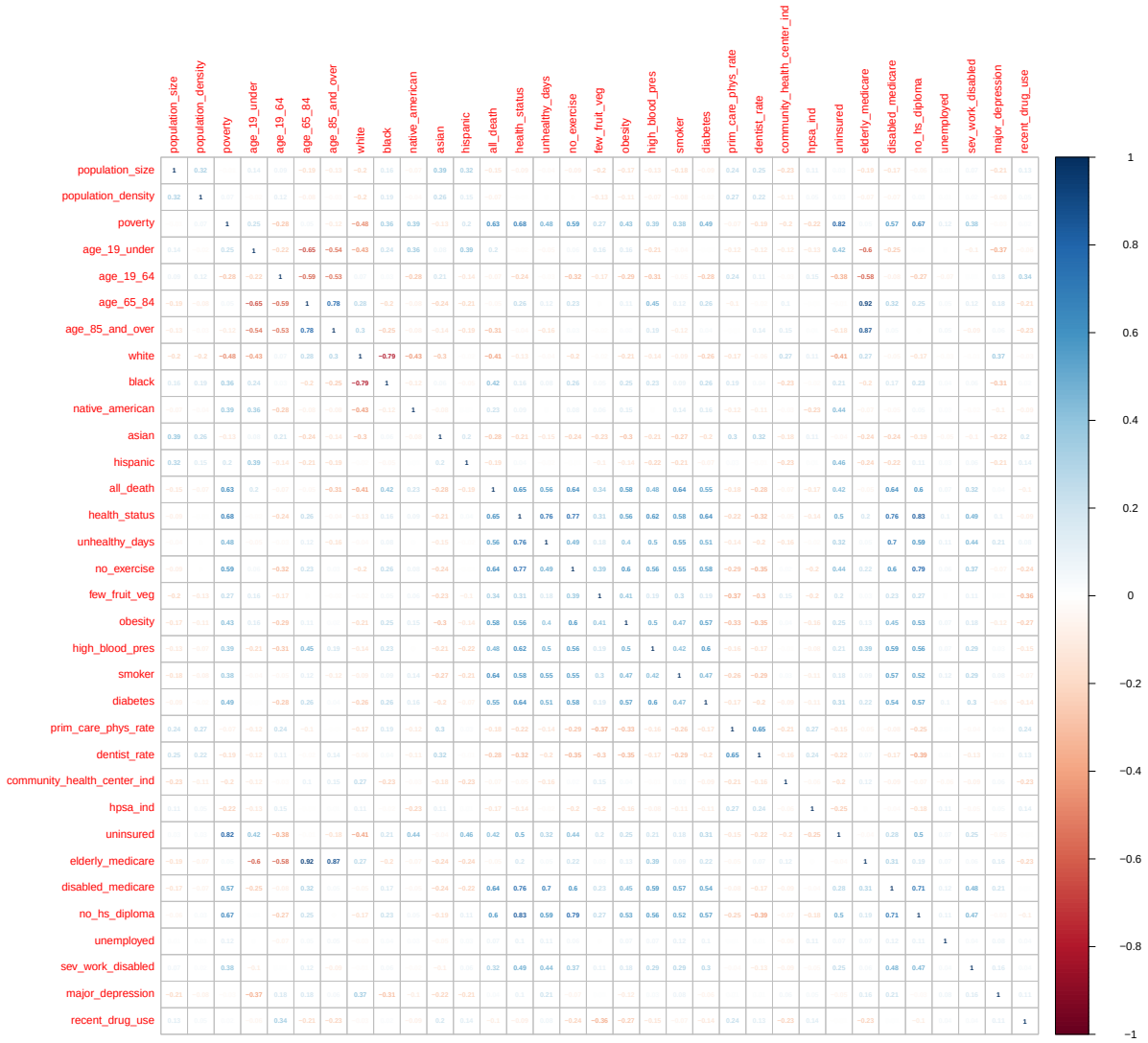

Figure 2: Hierarchical Clustering at the State Level

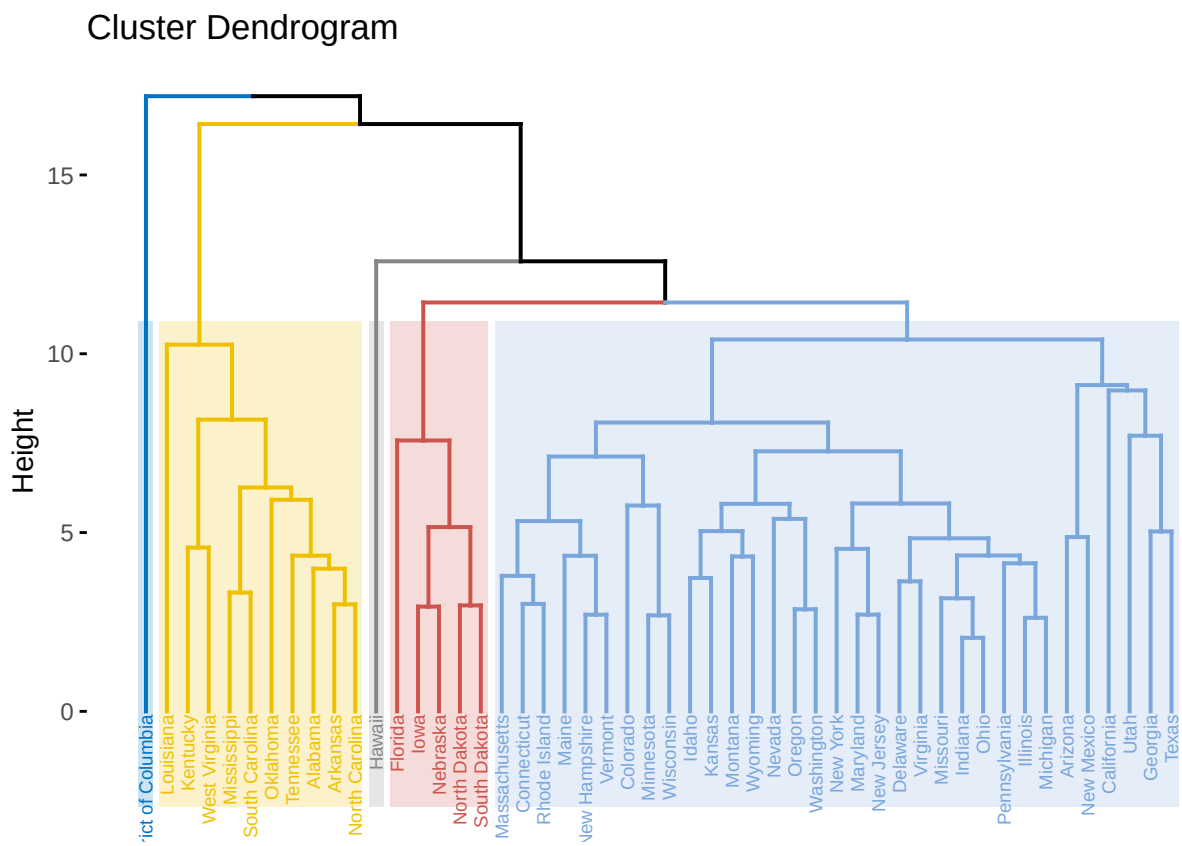

Figure 3: Heatmap at the State Level