STAT 663

Final Project

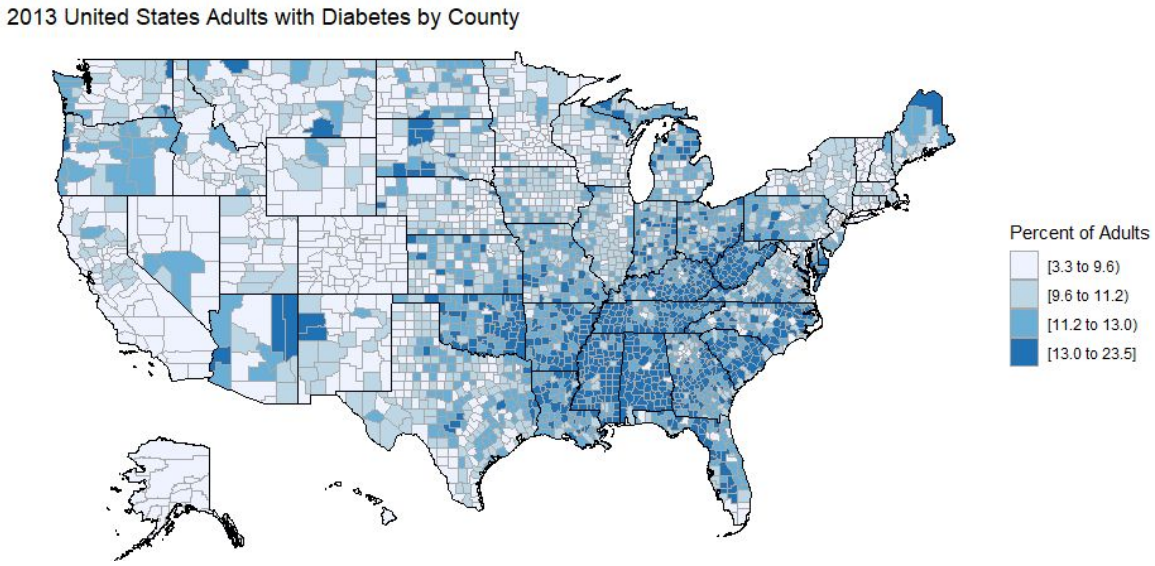
Group 5: Anais Jojic, Dan Cheng, Jiwan Hwang

[Dataset 1: Food Environment Atlas]

[Data Introduction and Exploration]

The first dataset we will be analyzing is the Food Environment Atlas (last updated March 27, 2018) from the United States Department of Agriculture found under the agriculture section of data.gov. It contains county level data for the entire United States. It has 3,143 rows of data, one for each of the 3,143 counties and county-equivalents. It has 277 indicators, including multiple years of the same measurement and counts, rates, and percentages of the same measurement.

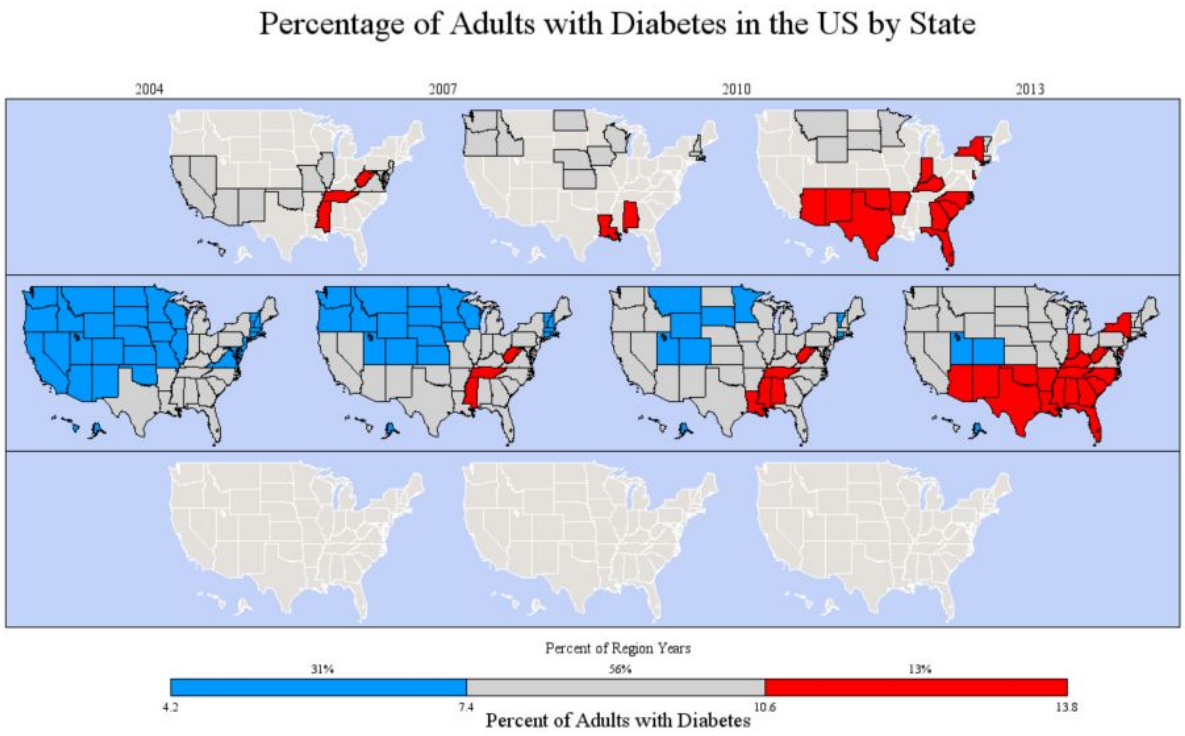
The indicators are broken up into nine categories titled access and proximity to grocery store, store availability, restaurant availability and expenditures, food assistance, state food insecurity, food prices and taxes, local foods, health and physical activity, and socioeconomic characteristics. The health category has a variable for the 2013 adult diabetes rate. This is the data point that was selected to be the dependent variable in our analysis.



<Choropleth Map of The Percentage Of Adults With Diabetes By County>

The choropleth map graphs the percentage of adults in the United States with diabetes, broken up by county. The percentages are broken up into four buckets so that viewers can easily distinguish between different counties and ranges without getting overwhelmed. The color pallet conforms with the ColorBrewer suggestion for sequential data, which is good for evaluating ordered data.This pallet will not cause issues for anyone with colorblindness. From the map we can see that the south east has a higher percentage of adults with diabetes than the northern and western states.

The Food Environment Atlas contains two years of data for the adult percentage rate of diabetes by county. We aim to get an idea of the historical pattern of diabetes in the United States over time. We will create a "rising and falling" map of the percentage of adults by state over multiple years. Here, we are splitting the county by states so that viewers can more easily discern patterns between the 50 states plus DC, rather than trying to compare 3000+ data points per map. The State of Obesity (stateofobesity.org/diabetes) has the percentage of adults with diabetes by state for all the years from 1990 to 2017.



<Rising and Falling Map of the Percentage of Adults With Diabetes By State Over Four Years>

Plotting the years from 2004 to 2013 by an increment of three years shows a clear pattern of increase in the percentage of adults with diabetes. Only the five states including Utah, Colorado, Michigan, Maine, and Pennsylvania did not move to a higher bucket over the course of ten years. The "rising and falling" map agrees with the choropleth map, showing even more clearly that the south and east of the US has a higher percentage of adults with diabetes than the northern and western states. However, the historical perspective shows that increased rates of diabetes is a problem for the entire United States.

[Experimental Lasso Regression Analysis]

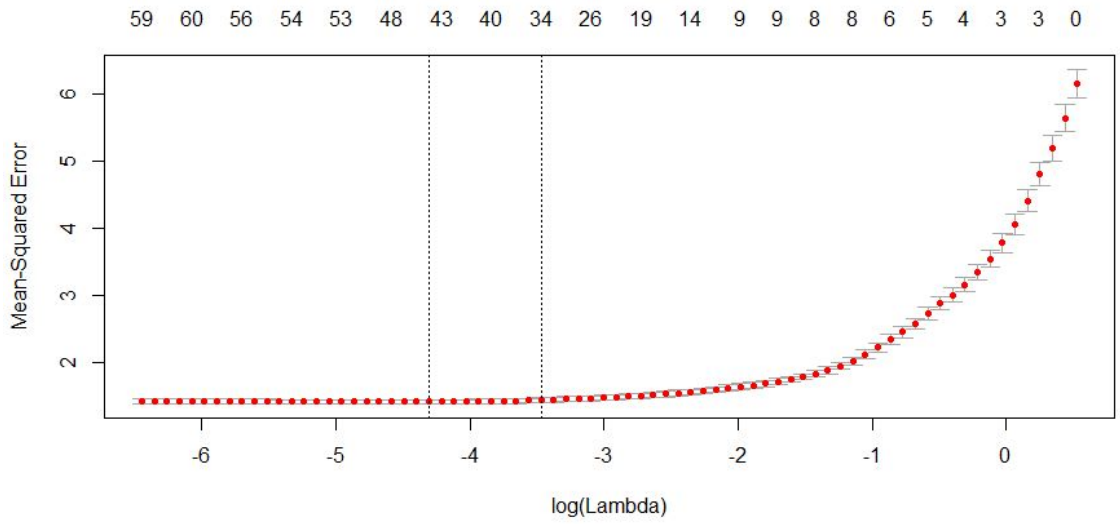
The first dataset for lasso regression was created by taking a subset of 78 variables from the original 277 indicators in the Food Environment Atlas. The original dataset had measurements listed multiple times, once as a count and again as a percentage or rate. The count variables were removed since percentages and rates are more conducive to comparing different counties. The dataset also had percent changes between the same measurement for two different years. These were removed since we are trying to predict the percentage of diabetes for the year 2013 and not in comparison to another time. These changes, along with a few removals for variables that had missing county data, led to the 78 variable dataset.

Lasso regression returns a model that uses a subset of the input variables, also known as a sparse model. It does this by having coefficient values of zero for some variables. The coefficient values depend on the value of the λ parameter. The goal of the first run with lasso regression is to obtain a smaller subset of the dataset, which will then be used to run a second round of lasso regression. The result from the second lasso regression should be significant as well as easily interpretable.

To obtain this smaller dataset we will look at the output results for the λ parameter with the smallest mean squared error. The cv.glmnet() function runs 10 fold cross validation to compute the average mean squared error and standard deviation for a range of λ. The λ with the smallest mean squared error is returned using lambda.min.

The graph below visually displays the mean squared error of the different λ models on the y-axis and the log(Lambda) values on the x-axis. The top row of numbers represents the number of variables in the λ models with non-zero coefficients. The vertical line crossing the

graph on the left is where the λ parameter that produces the smallest mean squared error is. The vertical line on the right is the largest value of λ with a mean squared error that is within one standard deviation from the minimum mean squared error and is returned using lambda.1se.



<Plot of Average Mean Squared Error and Standard Deviation As The λ Parameter Changes>

The value of the minimum λ is 0.013489. It has a mean squared error of 1.399 and a standard deviation of 0.0329. The min model keeps 45 out of the 78 variables in the dataset. The 1se model keep 34 variables out of the original 78. Since both models have too many variables to be able to interpret, we will consider the 45 variables kept in the min model.

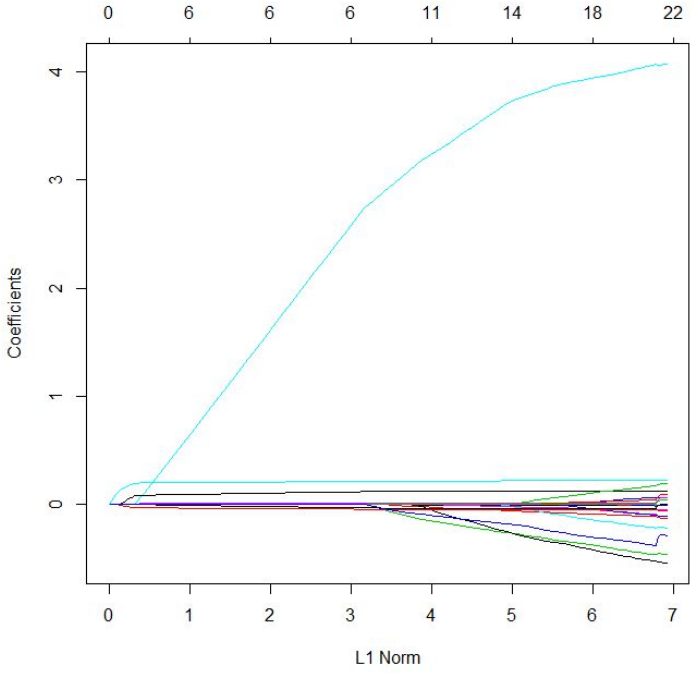
From the 45 variables, we will remove ones that are the same measurement for different years. For example, PCT_OBESE_ADULTS08 (the percentage of obese adults in 2008) and PCT_OBESE_ADULTS13 were both kept in the min and 1se model, but we will only consider PCT_OBESE_ADULTS13 for our final reduced dataset. After this clean up, we are left with a dataset of 23 variables.

[Final Lasso Regression Analysis]

To run lasso regression on the 23 variable dataset, we import it into R and split it into a training set and a testing set. The training set contains a random subset of 1572 rows and the testing set contains the other 1571 rows of the original 3143 counties and county-equivalents. 10

fold cross validation will be run on the training set to find the best λ parameter and the resulting lasso regression model will be evaluated using the testing set.

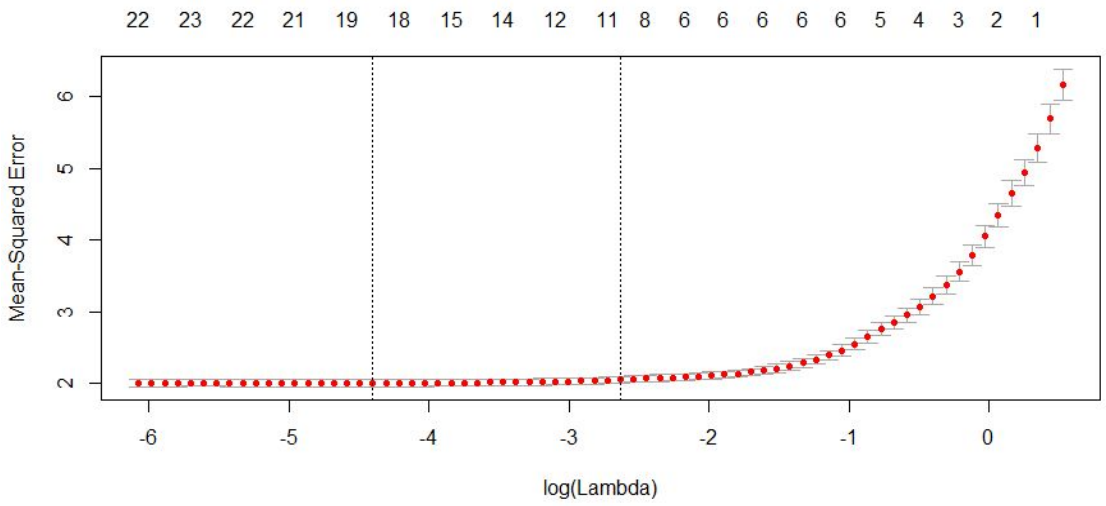
The following plot shows the changes to the lasso regression model as the λ parameter goes from large, forcing variable coefficients to zero, to small, allowing all 23 variables to stay in the model. The six variables with the largest absolute value coefficients in the full model are LN_PC_FFRSALES12, RECFACPTH09, FFRPTH09, PCT_NHPI10, PCT_WIC09, and PCT_OBESE_ADULTS13. We will see that five out of these six end up in our 1se model. PCT_OBESE_ADULTS13 is the last variable to stay in the model as all other variable coefficients go to zero.



<Plot of Model Coefficient Changes as L1 Norm Changes>

After running lasso regression using 10 fold cross validation on the training dataset, we get that the min lasso regression model has lambda value 0.01228. It has an average mean squared error of 2.010965, a standard deviation of 0.05051096, and keeps 17 of the 23 predictor variables. The min model can be seen by the left vertical line on the plot below. The 1se model is the model with the largest lambda value with an average mean squared error less than 2.0615

(the mean squared error of the min model plus its standard deviation). This lambda value is 0.07194. The 1se model uses 10 of the 23 available predictor variables.



<Plot of Average Mean Squared Error and Standard Deviation As The λ Parameter Changes>

After running the predictions for the lasso regression with lambda equal to the min and 1se values on the testing set, we get that the min lasso regression model has a mean squared error of 1.880619 and an R^2 score of 0.693. The 1se lasso regression model has a mean squared error of 1.923643 and an R^2 score of 0.686.

The two models are comparable in terms of how well they predict the dependent variable, but the 1se model is easier to interpret since it has seven less variables. The following table has the coefficients and variables for the 1se model.

Model Variables	Coefficients	Model Variables	Coefficients
(Intercept)	-11.81554	MEDHHINC15_1000s	-0.04504
PCT_LACCESS_LOWI10	-0.00025	PCT_NHBLACK10	0.01735
FFRPTH09	-0.16187	PCT_HISP10	-0.03072
LN_PC_FFRSALES12	2.63578	PCT_NHPI10	-0.07063
PCT_SNAP12	0.13177	PCT_OBESE_ADULTS13	0.21318
PCT_WIC09	-0.00071		

Four of the variables have positive coefficients that increase the predicted percentage of adults with diabetes per county. The other six variables have negative coefficients and the intercept is negative as well. The percent of obese adults in 2013 is closely related to the percent of diabetes based on the variable staying in the model the longest and the relatively large coefficient value.

[Lasso Regression Conclusion]

Lasso regression is a good model to use when you have a dataset with too many variables. For our dataset, it was instrumental in reducing the number of variables from 78 to 10. It performed variable selection by measuring the mean squared error for 100 different lambda values to identify the most influential variables. Among the resulting model options, we were able to select a decent linear model that is easier to interpret than the original full model without sacrificing significance.

[Introduction of Linear Regression Analysis]

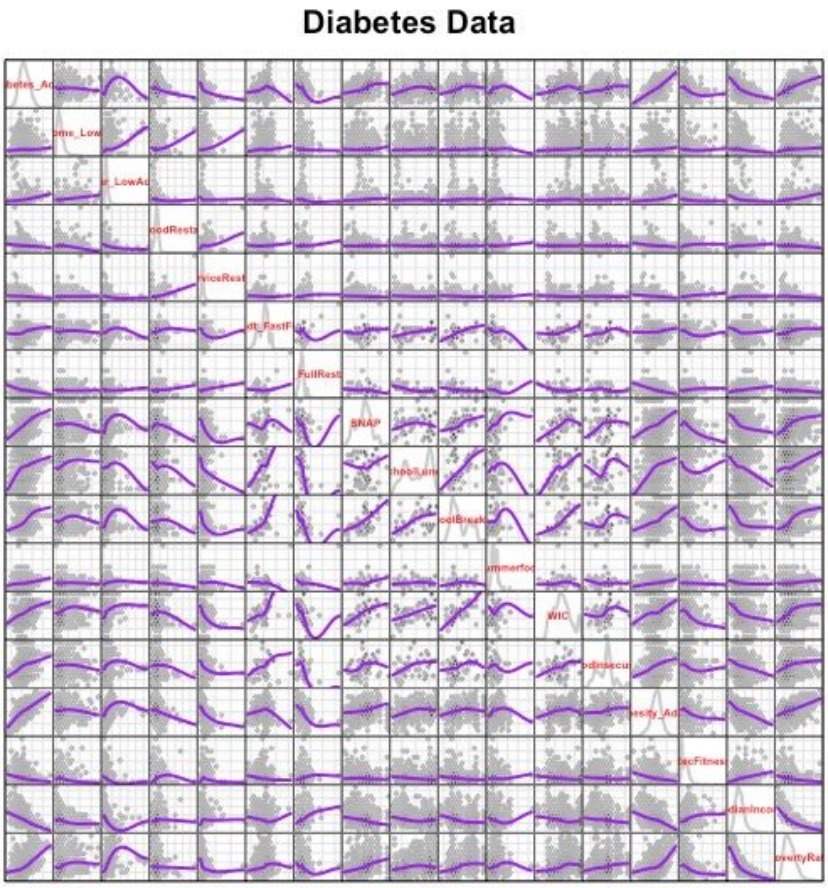
Linear regression is a statistical learning method for supervised learning. Supervised learning is the data mining algorithm using a function from labeled training data. After estimating the coefficients from the linear regression analysis, an equation can be written. This equation is the best fit to show relationships between the response variable and predictor variables, so the responses that can be expected from the model are always quantitative in nature. Having common knowledge about the relationship between variables makes it easier to build a linear regression model.

For the next step, we will explore the data. In addition, two approaches of subset selection: 10-fold cross-validation and best subset selection using model evaluation criteria were performed and compared with the pre-selected variables that consists of one response variable and seventeen predictor variables from the raw dataset.

[Data Exploration for Linear Regression]

18 variables from the original dataset were selected for new dataset to model a linear regression based on common knowledge (ex: relationship between obesity and diabetes), assumptions (ex: relationship between fast food and diabetes), and curiosities (ex: relationship between median income and diabetes). This new dataset contains only data of the latest years.

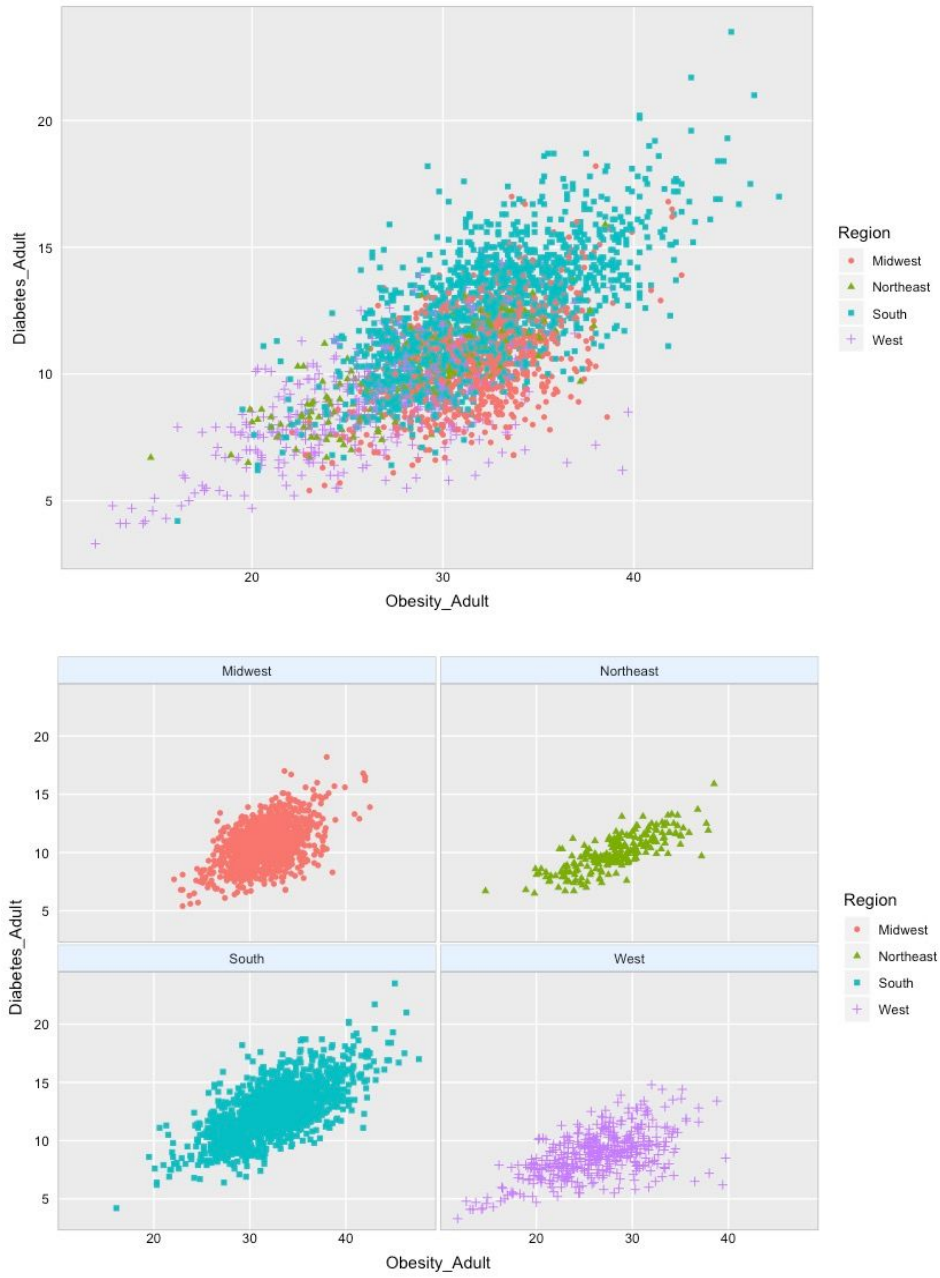
When analyzing a regression model, the scatterplot matrix is a great first step to scan any correlation between multiple variables. Each scatterplot shows the relationship between a pair of variables. We built a scatterplot matrix with hexagon binning and smoothes. Because the binning and smoothing functions for a scatterplot matrix cannot be performed for categorical variables, metro or non-metro counties, the only binary variable of 17 predictor variables, was removed. Based on our common knowledge about diabetes, the adult diabetes rate variable (Diabetes_Adult) is selected as dependent variable and the other the rest of 17 variables are independent variables.



<Scatterplot Matrix Full Numerical Variables Version With Smooth (Using Splom)>

On the plot above, the dependent variable is at the top left corner, so the first row can be scanned firstly to find any relationship between the dependent variable and other independent variables. For example, the relationship between the dependent variable, adult diabetes rate and another independent variable, adult obesity rate (14th column on the scatterplot matrix) have a very strong positive correlation.

The plots below result from performing a scatterplot of the strong positive correlation between adult diabetes rate and adult obesity rate and its faceting process in terms of U.S. regions defined by Census Bureau. From these plots, South has the bigger adult obesity and diabetes problems than other regions.

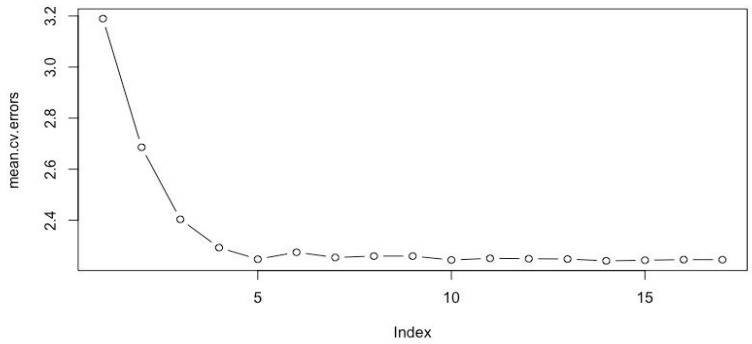


<Scatterplot and Faceting>

[Subset Selection Using 10-fold Cross Validation]

First, the subset selection using 10-fold cross validation was performed. From the result, the average Mean Squared Error (MSE) of 10-fold Cross Validation (CV) for model 14 (2.240978) is the lowest among the seventeen models. However, seeing the plot below, from model 7 to model 17, the averages of 10-fold MSE has no large differences (Max: 2.259660 and

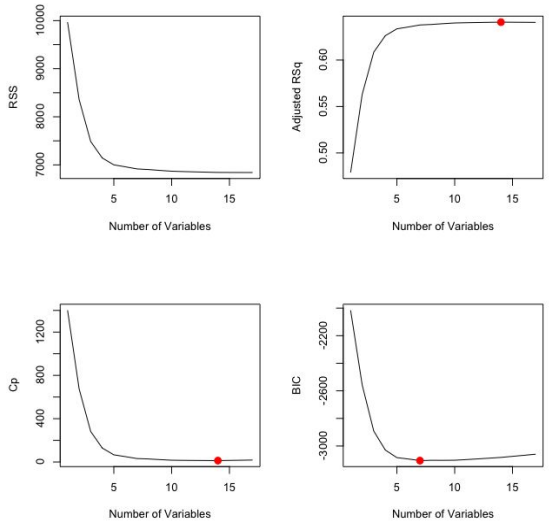
Min: 2.240978). To compare the subset selection using 10-fold CV and another approach, other subset selection using model evaluation criteria was performed next.



<Mean Squared Errors: 10-fold Cross Validation>

[Subset Selection Using Model Evaluation Criteria]

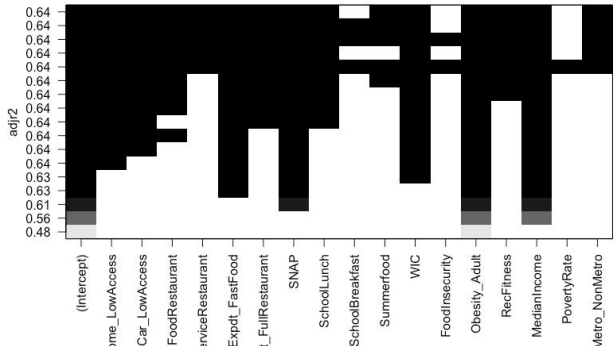
Second, the subset selection using model evaluation criteria was performed. This approach is more traditional method for subset selection. Seeing the table below, the red dots show the number of variables that is best to use in terms of each of the three criterion: Adjusted R-squared, CP, and BIC. In terms of both Adjusted R-squared and Mallows's Cp-statistic, using the best 14-variable has the largest Adjusted R-squared value (0.640802) and smallest Mallows's Cp-statistic (12.75998). For BIC, using the best 7-variable results in the smallest BIC value (-3106.013).



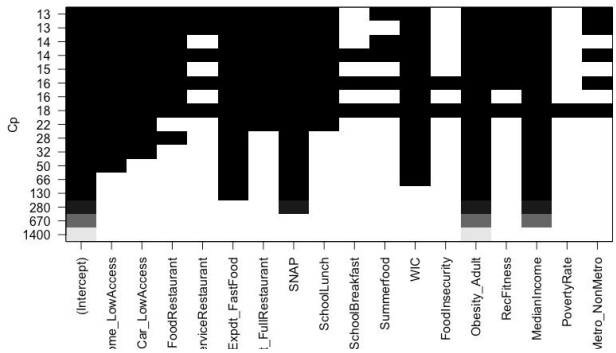
<Model Evaluation Criteria>

[Panel Plots with Variables (Adjusted R-squared, Cp, BIC)]

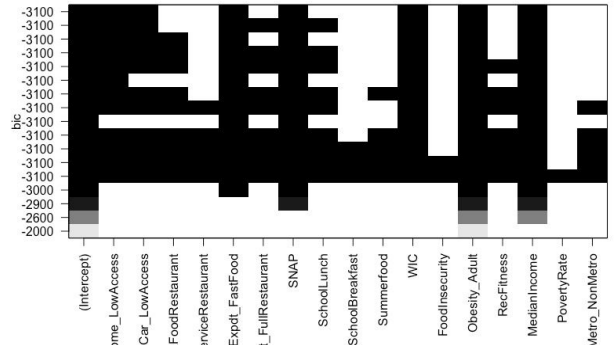
Using panel plots, it is easy to see which variables and how many variables are selected in each best subset selection model. The best model is at the top line of each panel, and the worst is at the bottom. For example, for adjusted R squared, the 14 variables among 17 variables are chosen for the best subset selection model. 7 variables are chosen for the best subset model in terms of BIC.



<Adjusted R-squared>



<Mallows's Cp-statistic>



<BIC>

[Conclusion from Two Subset Selection Approaches]

Even though the BIC is often chosen as the preferred model selection criterion, for this project, the best 14-variable model was chosen. This is because all three results of 10-fold CV and the other model selection criterion, Adjusted R-squared and Mallos's Cp, show the same number of variables, which is 14. Moreover, the MSE of the best 14-variable model is 2.196737 that is lower than the MSE of the best 7-variable model, 2.220179 (but there is still no large differences between them). Summary of the best 14-variables model and their coefficients are in table below.

Best 14-variable	Coefficients	Pr(>|t|)	Best 14-variable	Coefficients	Pr(>|t|)
(Intercept)	1.899	0.000370	Schoollunch (% pop, 2015)	-0.079	0.000449
LowIncome_LowAccess (to store, %, 2015)	-0.019	5.71e-08	SummerFood (% pop, 2015)	0.135	0.110690
NoCar_LowAccess (to store, %, 2015)	0.040	0.000166	WIC (% pop, 2015)	-0.534	4.62e-10
FastFoodRestaurant (/1,000 pop, 2014)	-0.296	0.003384	Obesity_adult (rate, 2013)	0.268	< 2e-16
FullServiceRestaurant (/1,000 pop, 2014)	0.107	0.056731	Rec and Fitness (/1,000 pop,2014)	-0.808	0.051315
Expenditure FastFood (per capita, 2012)	0.007	< 2e-16	MedianIncome (/1,000 pop, 2015)	-0.050	< 2e-16
Expenditure FullRest (per capita, 2012)	-0.001	0.000166	Metro_Nonmetro (counties, 2010)	0.107	0.096992
SNAP (% population, 2016)	0.164	< 2e-16	**Variable names changed for convenience**		

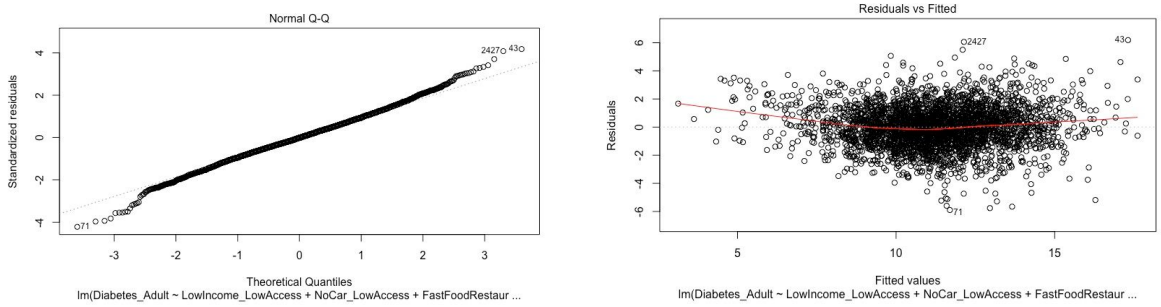
<TABLE: Results from the Best 14-Variables Model>

Except for four variables, the other 10 variables are statistically significant, and the p-value of the linear regression model using the best 14-variable (< 2.2e-16) is also significant. The adjusted R-squared is 64.08%.

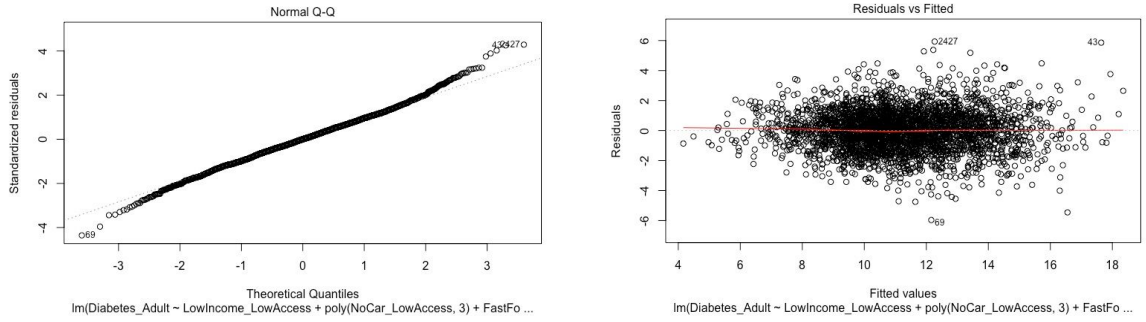
[Diagnostic Plots: QQ-plot and Residual vs Fitted plot]

For the last step of linear regression analysis, there are four kinds of diagnostic plots which are QQ-plot, Residual vs Fitted plot, Scale-Location plot, and Residual vs Leverage plot.

For this project, our team analyzed QQ-plot and Residual vs Fitted plot. The first pair of the plots below results from the linear model using the best 14-variable. The second pair of the plots results from another linear model with polynomial terms based on the best 14-variable. The variables transformed to polynomial terms are selected by scanning the scatterplot matrix. This model has an improved adjusted R-squared (68.74%) with same statistical significance (< 2.2e-16). Comparing the QQ-plots, the model with polynomial terms has similar normality. Looking at the Residuals vs Fitted plot, the red line, which is a scatterplot smoother, is flatter and more horizontal. It shows that there is no discernible non-linear trend to the residuals.



<Diagnostic Plots of the Best 14-Variable Model>



<Diagnostic Plots of the Best 14-Variable Model with Polynomial Terms>

[Dataset 2: US Lung Cancer]

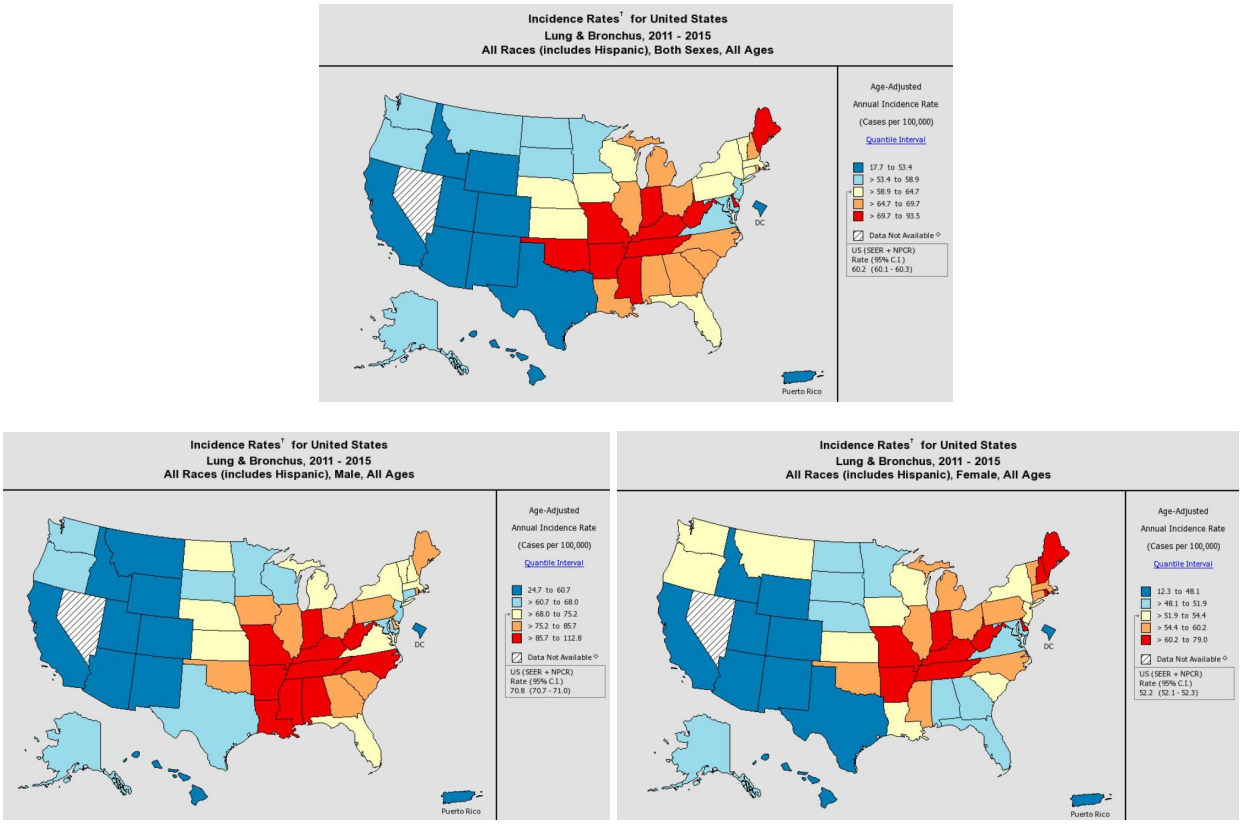
[Data Introduction]

Cancer is a severe public health problem in most of the countries around the world. Though medical technology develops quickly, it is still one of the most common causes of death. According to American Cancer Society, the overall cancer mortality rose over the past decades and lung cancer death rate keeps increasing rapidly due to the tobacco use (Siegel et al, 2017). The data from SEER (https://seer.cancer.gov/statfacts/html/lungb.html) also shows that the percentage of surviving in 5 years is only 18.6%, which is very low.

Therefore, the second dataset we selected is the US lung cancer dataset by states, which is also related to the health topic as our first dataset is. Comparing with the first dataset, our second dataset is relatively small since it only includes 18 variables and 50 records for each state. The data are combined by tobacco legislation information from Centers for Disease Control and Prevention, current smoking data from National Cancer Institute, state cancer profiles (including incidence rate by race and by sex, cancer screening information, education, income) from National Institute of Health, and demographic statistics (population) from US Census Bureau.

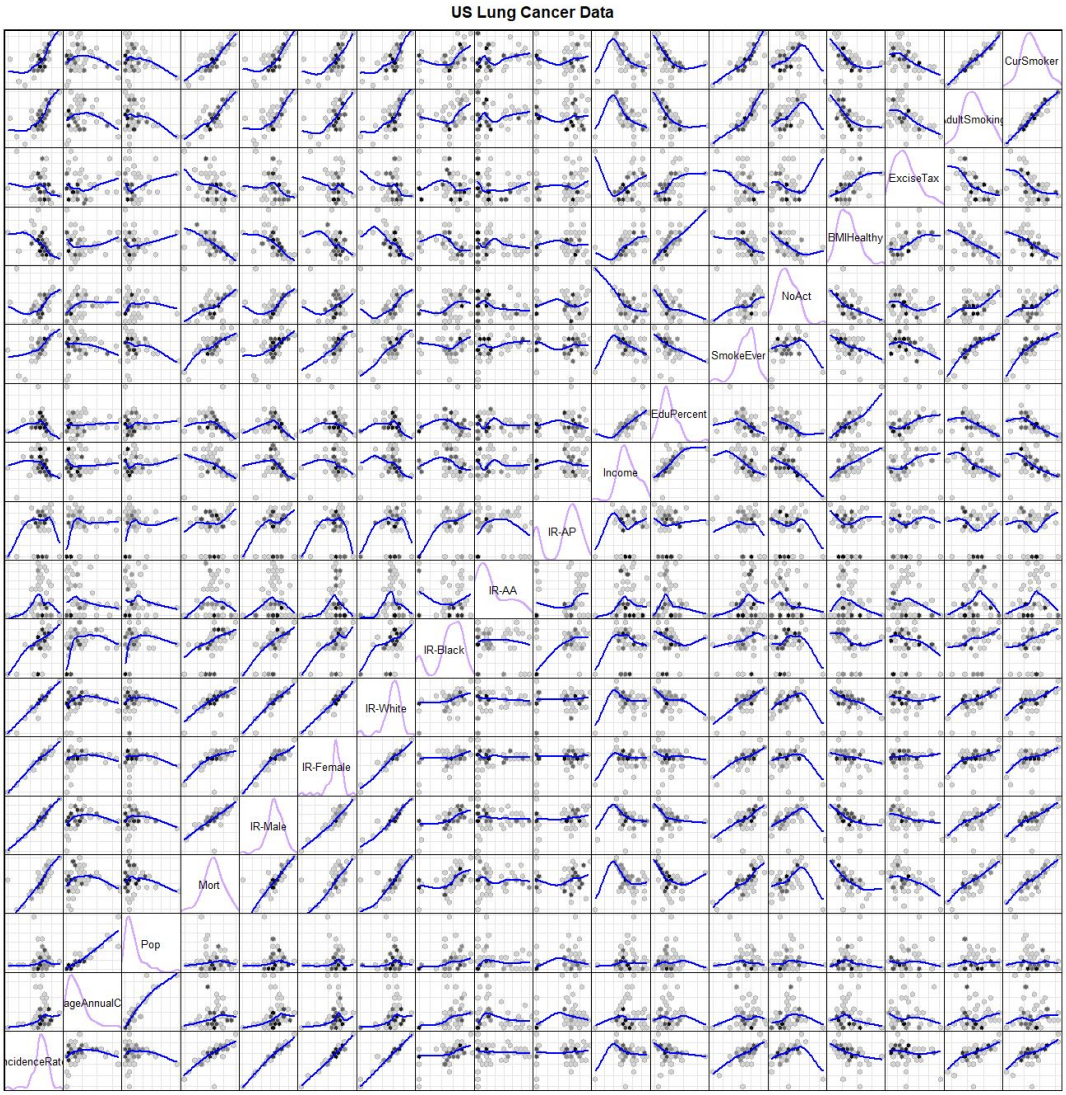
[Data Exploration]

According to the Lung Cancer Fact Sheet from American Lung Association, lung cancer has significant gender and race differences. There are more men diagnosed with lung cancer each year and the death rate of African American is much higher than any other racial or ethnic group. Therefore, we will first explore the spatial patterns of incidence rate between different genders and races. The three maps below are the incidence rates of lung cancer for both sexes, only male, and only female in United States. These choropleth maps are generated from the interactive map application designed by National Institute of Health (https://statecancerprofiles.cancer.gov/map/). As we can see, the spatial patterns of the incidence rate for male and female are quite different since the rate value of female is much smaller than male and the area of hotspot states for female are less than the area of male incidences. Due to the data limitation, the records of Nevada are not shown in the maps.



< Choropleth Maps: The Incidence Rate of Lung Cancer for Both Sexes, Only Male, And Only Female>

The scatterplot matrix is an efficient way to scan and find any correlation between variables. So in the second step of data exploration, we plot the scatterplot matrix with hexagon binning and smoothes for all variables in order to get an overview of the potential relationship between every two variables. We choose “Incidence Rate” as the dependent variable and the other 17 variables as independent variables. Since our dependent variable is “Incidence Rate”, we can focus on the scatterplots in the last row of the matrix.

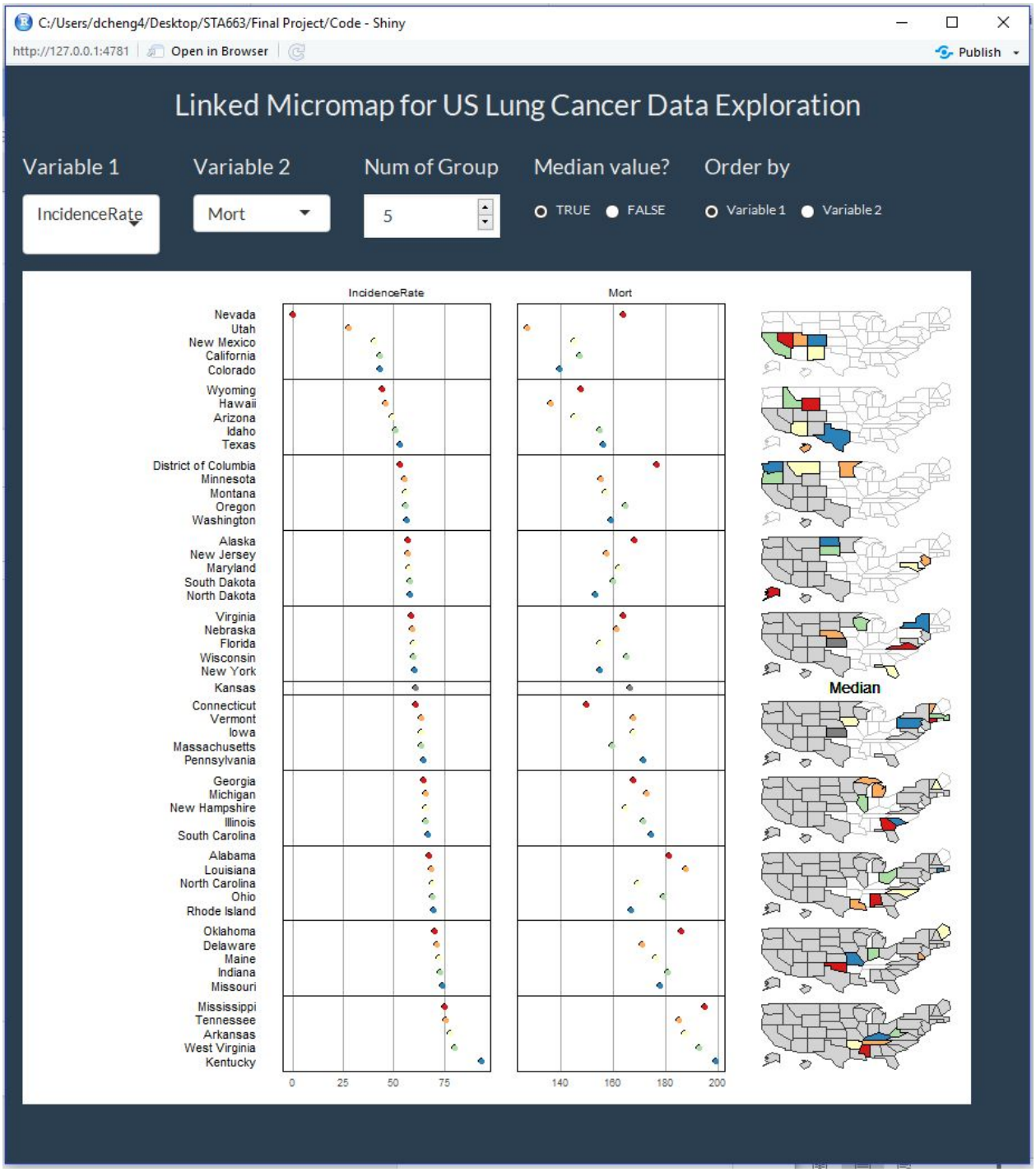


<Scatterplot Matrix for All Variables.>

Although the view of each scatterplot is very small, we can still find trends between incidence rate and other variables. For example, the incidence rate for both sexes is associated with mortality, which makes sense since the death rate is very much related to the number of people that were diagnosed with lung cancer.

Thirdly, we designed an interactive linked micromap application for US lung cancer data using two R packages, shiny and linkedmicromap. The shiny theme we chose is called "superhero", which has a dark blue background and is very helpful to highlight the map information. In the application, we designed to show dot charts of two variables at the same time so that people could easily identify the potential similar spatial patterns between two variables.

The linked micromaps are also shown along with the two variables we chose. As we can see from the figure below, variable 2 (Mortality) has a roughly similar trend with variable 1 (Incidence Rate) and the states located in the southeast area are likely to have a higher rate.

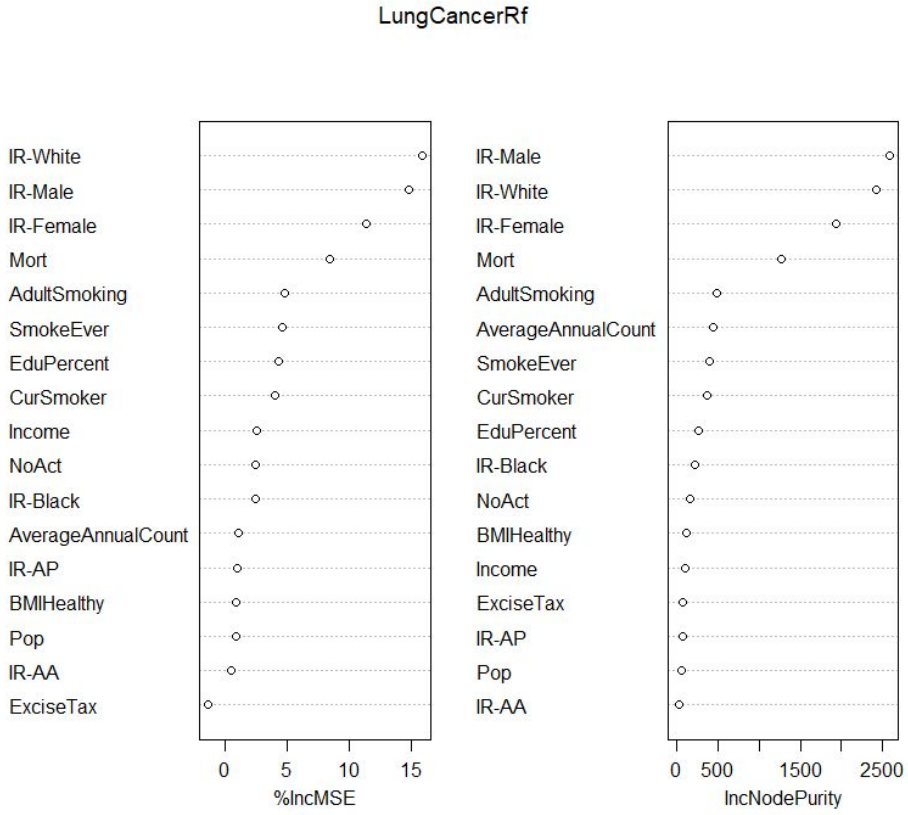


<Linked Micromaps for US Lung Cancer Data Exploration.>

[Random Forest Regression]

Our goal is to model the incidence rate of lung cancer in the United States using the other 17 variables. We chose to use Random Forest method to run the regression because it works well on find the important variable sets. There are two specification methods, one is the matrix column and the other is the formula method. In this study, we use the matrix column method to clarify the dependent variable and independent variables.

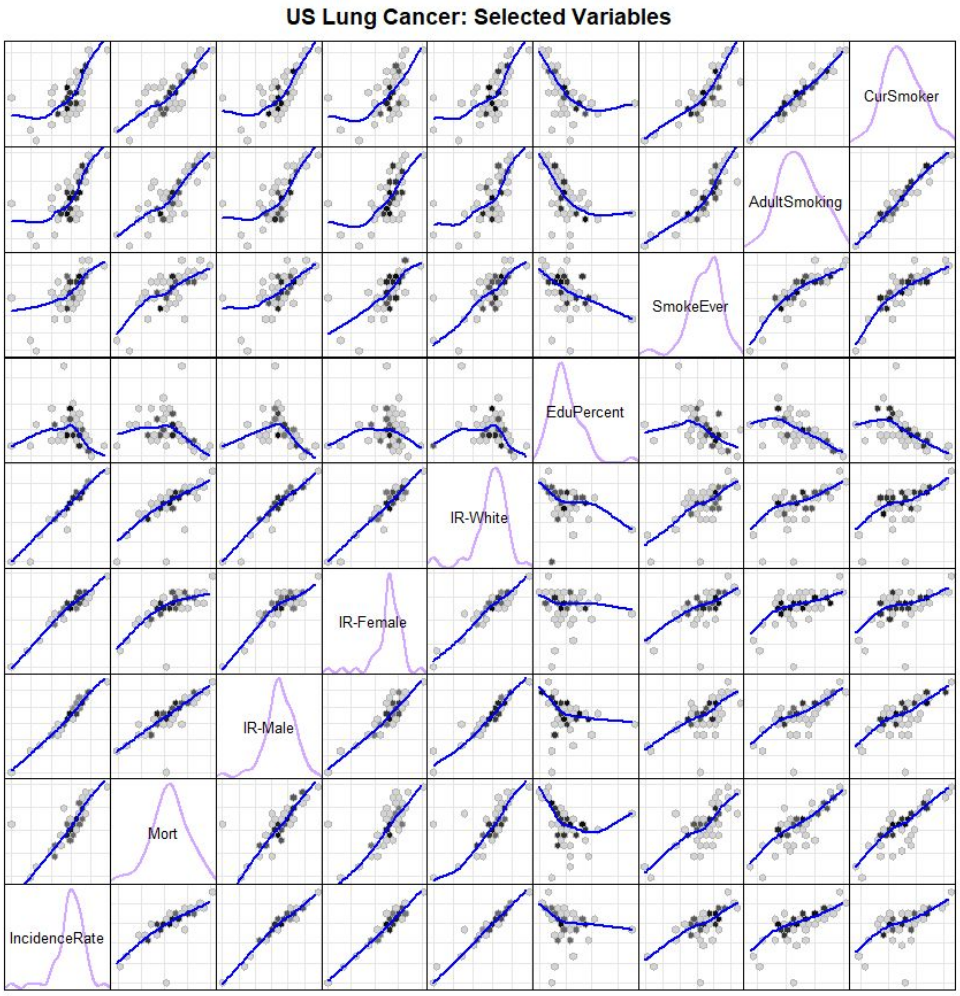
Firstly, we ran the random forest regression for all variables and draw the variable importance plot. The number of trees is set for 500 and the number of variables tried at each split is 5. The mean of squared residuals is 47.75218 and the percentage variable explained is 79.16.



<Variable Importance Plot>

Based on "%IncMSE" result, we first eliminate 5 lowest importance variables (ExciseTax, IR-AA, Pop, BMIHealthy, IR-AP) and then eliminate additional 4 lowest importance variables (AverageAnnualCount, IR-Black, NoAct, Income). After eliminate the first set of variables, the number of variables tried at each split decreases to 4. The mean of squared

residuals drops to 46.07061 and the percentage variable explained increases to 79.9. Then we continued to eliminate the 4 variables, the number of variables tried at each split decreases to 2. The mean of squared residuals drops to 41.65026 and the percentage variable explained increases to 81.82. The result is much better. So we stop to shrink variables and draw another scatterplot matrix for 8 selected variables.



< Scatterplot Matrix for 8 Selected Variables >

The scatterplot matrix above is for the 8 selected variables after removed the 9 consensus lowest importance variables. So the panels look larger and show a much clearer view of each scatterplot. As we can see from the scatterplot matrix, except for the Education Percentage, all other 7 variables are statistically significant and show an obvious increasing trend associated with the incidence rate for both sexes at all races.

[Conclusion]

Although our second dataset is small, we can still show that regression tree is easy and obvious to explain the data to people. The variable importance plot is very helpful to pick the most important variables. For the data exploration, we used choropleth maps, scatterplot matrix, and interactive linked micromap visualization using shiny. These three ways show a very clear view of the data and the potential association between variables. We believe it would contribute to not only the lung cancer issue, but also more health related problems in the future.

[Reference]

Economic Research Service. "Food Environment Atlas". United States Department of Agriculture. Accessed December 14, 2018. https://www.ers.usda.gov/data-products/food-environment-atlas/

The State of Obesity, https://stateofobesity.org/diabetes/

James, G., Witten, D., Hastie, T., Tibshirani, R. (2017) An Introduction to Statistical Learning with Applications in R. New York: Springer. (Original work published in 2013).

ColorBrewer 2.0, http://colorbrewer2.org/#

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. *CA: a cancer journal for clinicians*, *67*(1), 7-30.

https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html