Food Atlas Dataset

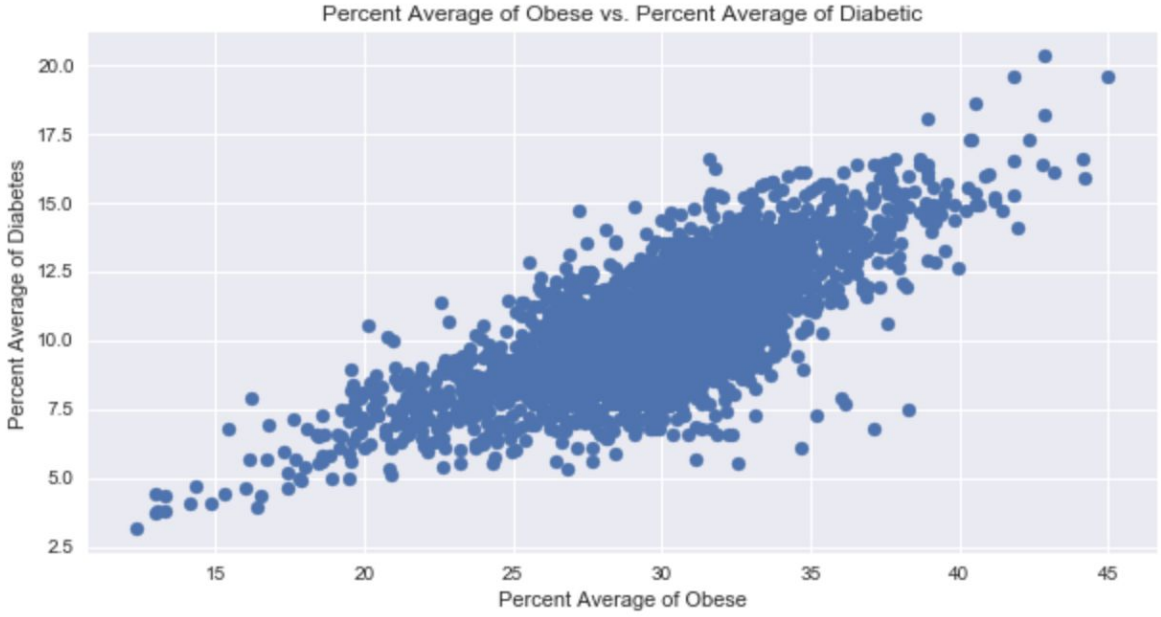
Data Cleaning and Visualization

For our project, we chose to use the Food Atlas Environment data to see if we can find any relationships between health and environmental and socioeconomic factors such as race and income. The data itself came in many sheets for many categories. We focused on accessibility, restaurants, stores, socioeconomic factors, and health statistics given to us. We hypothesized that an increasing number of fast food restaurants and higher income would have a linear relationship with the percentage of people with obesity.

To clean our data, we first look for missing values. According to the dataset, missing values came in the form of blank cells and cells with value "-9999". We were able to visualize this using the *missingo* library, and proceeded to remove all missing values. After that, we were able to pull out some fun facts about our dataset. Across the nation the percentage of racially white people is 78%. 15% of people are over the age of 65, 23% of people are age 18 and under, and thus we extrapolate that the 62% of the population are between 19 and 64. On average, 29% of the population are obese, significantly higher than the 11% who are diabetic.

To organize our data into one dataframe, we tried to take as much as we could and simplify. The dataset gives percentages of various factors such as obesity, income, diabetes, and number of grocery stores, for example, in various years. We decided to add the values associated with that year and take the average, so we can get one value of the variable we are interested in.

To plot numerical plots, we plotted something simply such as the average percentage of obesity versus the average percentage of diabetes on a scatter plot.

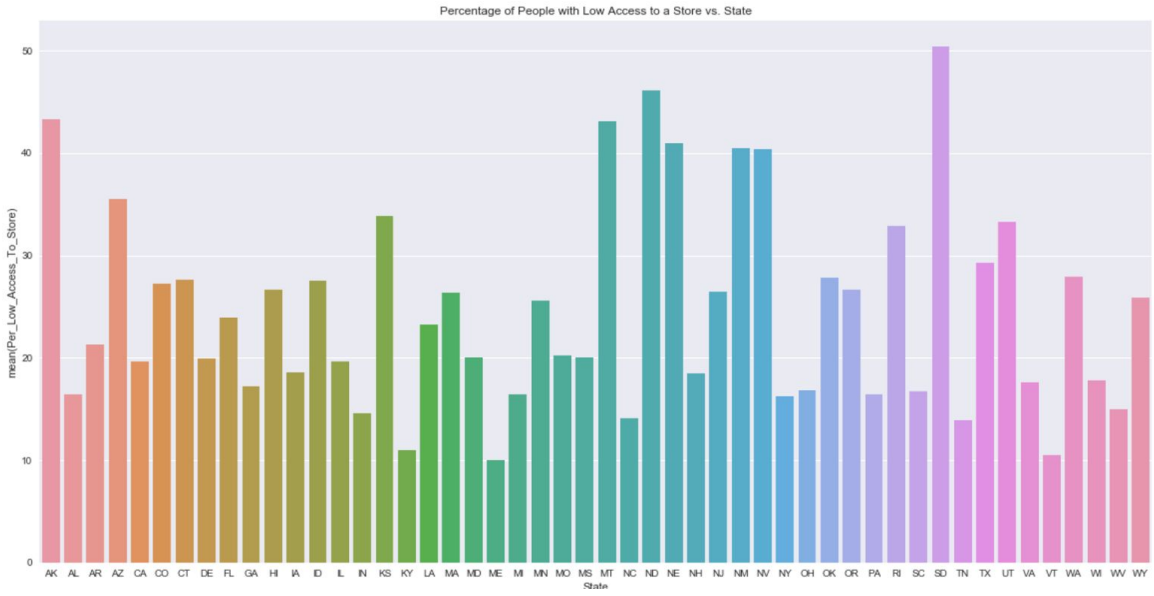


Here, it is clear that there is some relationship between obesity and diabetes. The scatter plot is plotting the average percentage of obesity within a county against the average percentage of diabetes within a county. There are about 3,100 counties.

Another variable we were interested in was the median household income. The plot above is one that compares the median household income versus the average percentage of obesity within all the counties. We took the log of the median household income, because it looked to fit more linearly. In any case, we saw a linear relationship between household income and obesity; as in, the higher household income, we saw the average obesity rate decrease.

We noticed that for this data, there were not many categories, so the only categorical plots we were able to make were against states, so it was difficult to make a story and categorize in a finer way. We grouped the data by states and plotted a violin plot to see the distribution of the percentage of obese people within the state.

From this plot, it looks like most states have a normal distribution of obesity within all the counties. The strange one is DC because it had only one county. Also, we see that Hawaii has a distribution that skewed, probably due to the low number of counties.



From this plot, we can extrapolate the the percentage of the state whose population have low access to stores such as grocery stores and super stores.

Linear Regression

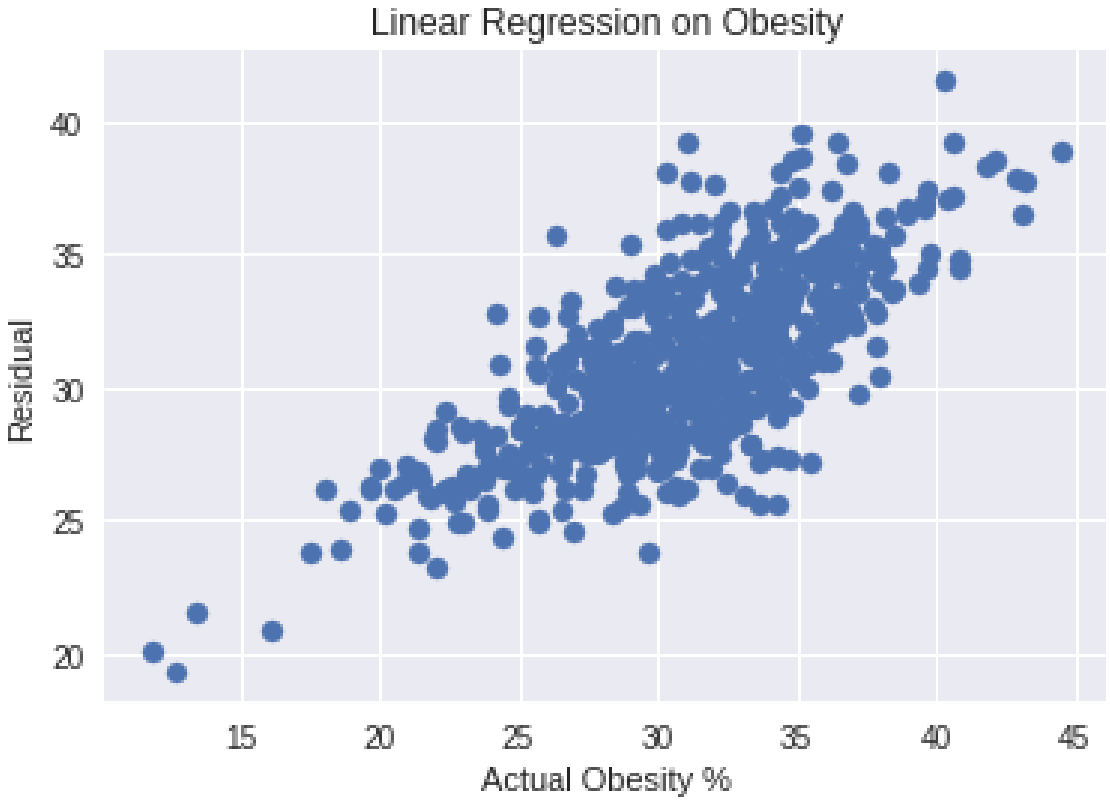
Introduction:

The goal of this linear regression is to not only to derive a reasonable predictor of obesity rate from other factors, but also see the exact relationship between these factors and obesity. As such, model interpretability is key. Thus, rather than regressing upon a wide range of variables, we instead plotted various single-relationship scatter plots between obesity and factors to see if there was a meaningful relationship.

As such, we eventually decided on the following variables for our first trial: White Population % 2010, Median Household Income 2015, Recreational Facilities per 1000 Population 2014, Adult Diabetes % 2013, Fast Food Stores per 1000 Population 2014, Grocery Stores per 1000 Population 2014, and Low Access % 2015. Though this data is scattered

across a range of 5 years in predicting the Adult Obesity % 2013, this is a limitation of the dataset and further qualitative analysis on these variables show insubstantial variation throughout the dataset. We grouped among counties, giving us 2674 data points after data cleaning.

Iteration 1:



Using SKLearn Linear Regression package, we divided the dataset into 80% training, 20% testing using a systematic random sample. Our initial results were:

MSE: 9.942

R^2: 0.492

Coefficients: 3.59832000e-03, -3.31024515e-05, -3.46575637e+00,

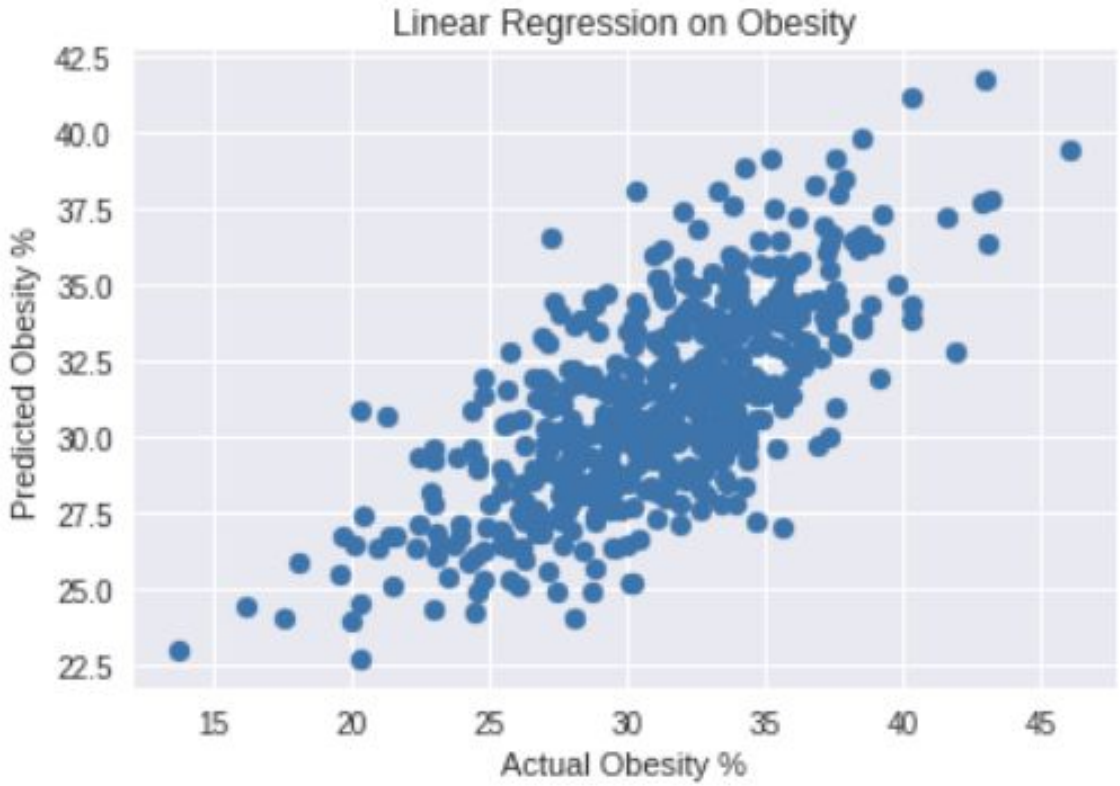
1.12599042e+00, -1.37364212e+00, -1.67660249e+00,

1.14270182e-03.

We immediately realize that the coefficients for White Population % 2010 and Low Access % 2015 were comparatively low and significant, constituting little more than noise for our model. Thus, for model interpretability, we will remove them for trial 2.

Note: though at first glance Median Household Income 2015 also has a low coefficient, its values range in the 10,000 while the others are %. After accounting for scale, it actually has a huge factor on our linear model. However, again for model interpretability, we will take the log of Median Household Income 2015 so its coefficient is more comparable to the others.

Iteration 2:



Running our second model with the above changes gives us the following results:

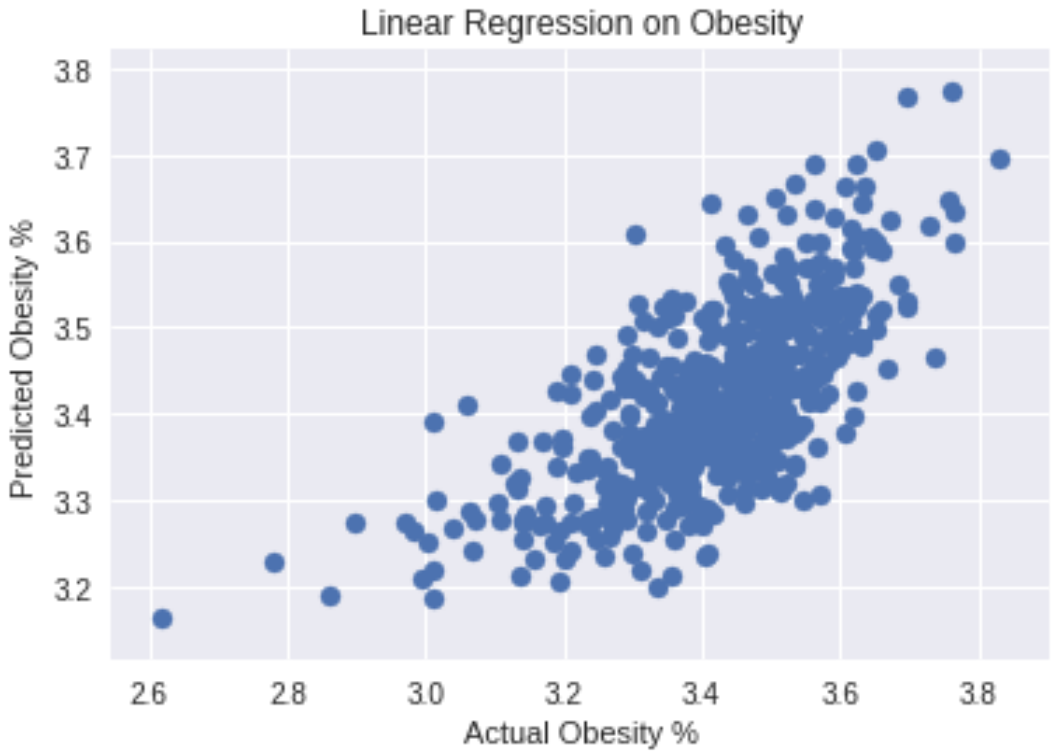
MSE: 9.934

R^2: 0.492

Coefficients: -1.63475014, -4.63242065, 1.09095174, -0.47499969, -2.44060587

These results were marginally better than Iteration 1 with a slightly lower MSE, but more importantly it supported our claims that the two dropped variables had negligible influence. Now, we will continue to optimize our model, though all changes only have minor impacts on MSE and R^2.

Final Iteration:



MSE: 9.567

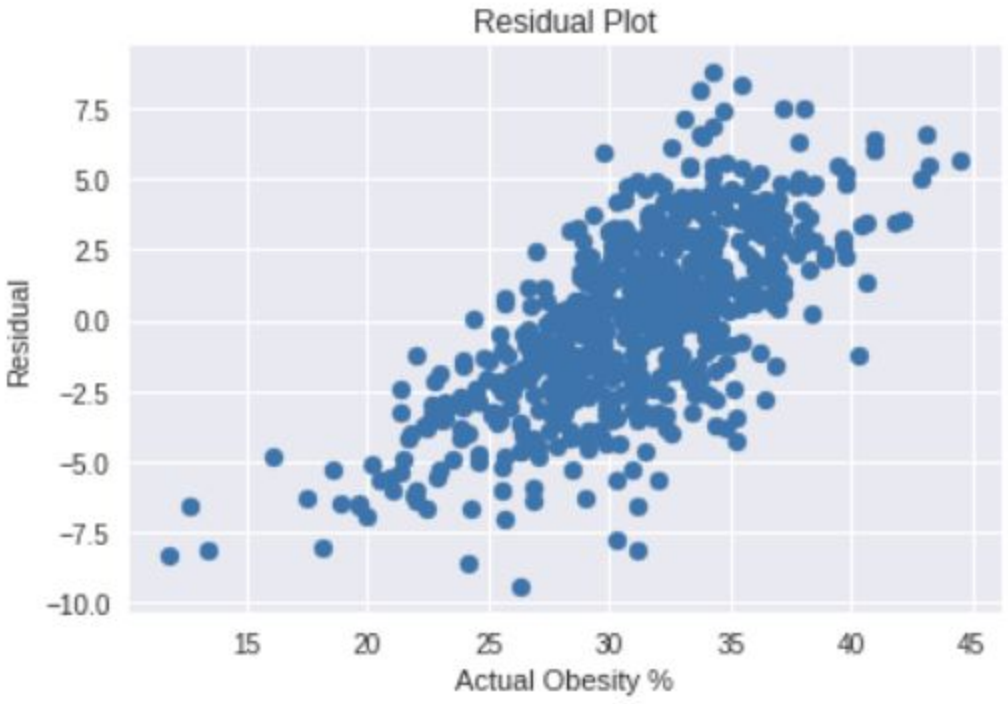
R^2: 0.526

Coefficients: -1.02919129, -4.68871626, 1.15794923, -1.26265804, -2.4274053

These are the final values for our linear regression model with the following parameters: log(Median Household Income 2015), Recreational Facilities per 1000 Population 2014, Adult Diabetes % 2013, Fast Food Stores per 1000 Population 2014, and Grocery Stores per 1000 Population 2014. It was achieved with a 75% train, 25% test split.

Analysis and Discussion:

Generally, a high R^2 value close to 1 implies that a linear fit is a good model for the data. However, datasets involving humans are intrinsically less predictable, and so we take our R^2 value of 0.526 as sufficient to support a good model. The more troubling information comes from the residual plot, where there is a clear positive relationship where we would normally expect a random/no relationship.



This relationship was present in all iterations of our linear model, and even after fitting the various values to log, x^2, x^0.5, etc, there was no way to “fix” our model. It is important to

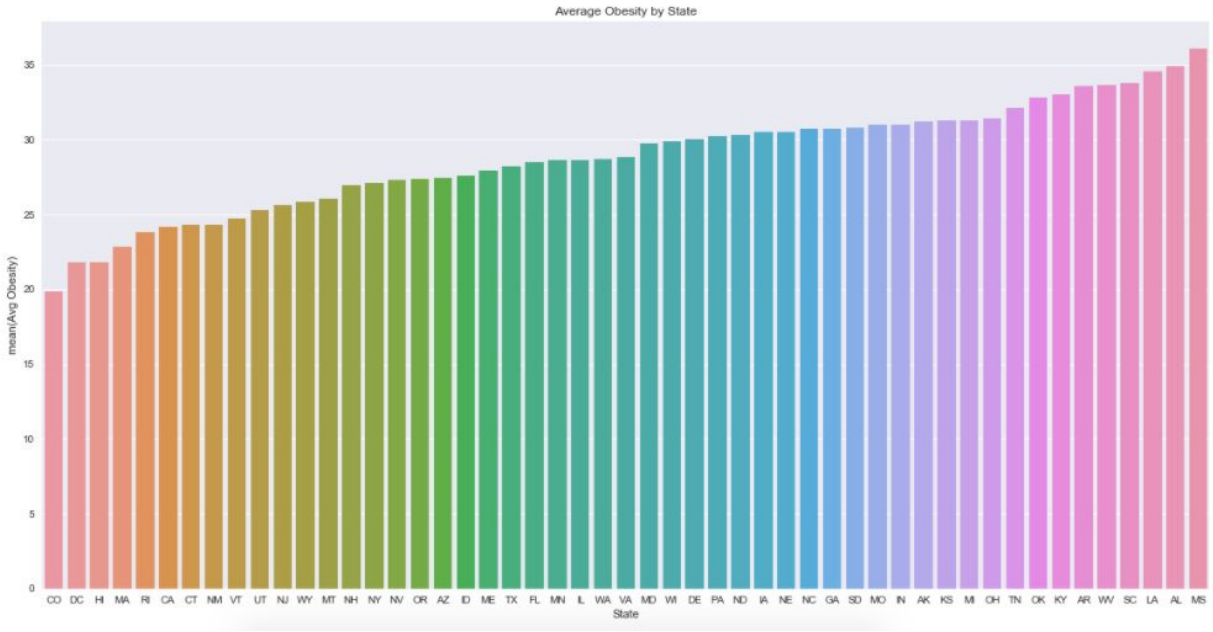
remember that multiple linear regression assumes the various variables are completely independent, yet this is obviously untrue in our case as the various factors definitely have confounding factors. Thus, though troubled by the trend in our residual plot, we accept it as a limitation of our dataset/goal.

Our final model has positive coefficients for Adult Diabetes % 2013, and negative coefficients for Recreational Facilities per 1000 Population 2014, Fast Food Stores per 1000 Population 2014, and Grocery Stores per 1000 Population 2014. Most of these agree with "common sense" as wealthier households normally imply more and healthier options, more gyms imply more exercise, a higher diabetes rate implies a less healthy population, and more grocery stores imply more local and fresh food options. However, the negative coefficient for Fast Food Stores per 1000 Population 2014 is surprising, as one would normally expect Fast Food Stores to be associated with obesity but our model implies a negative relationship between the two. *This could be a point of further analysis or research in the future.*

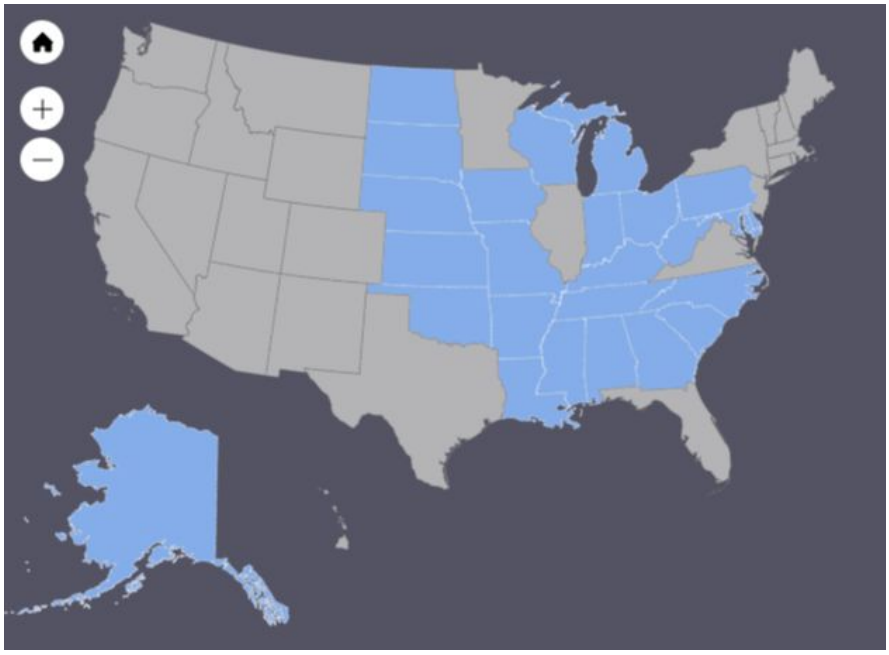
Logistic Regression

Introduction:

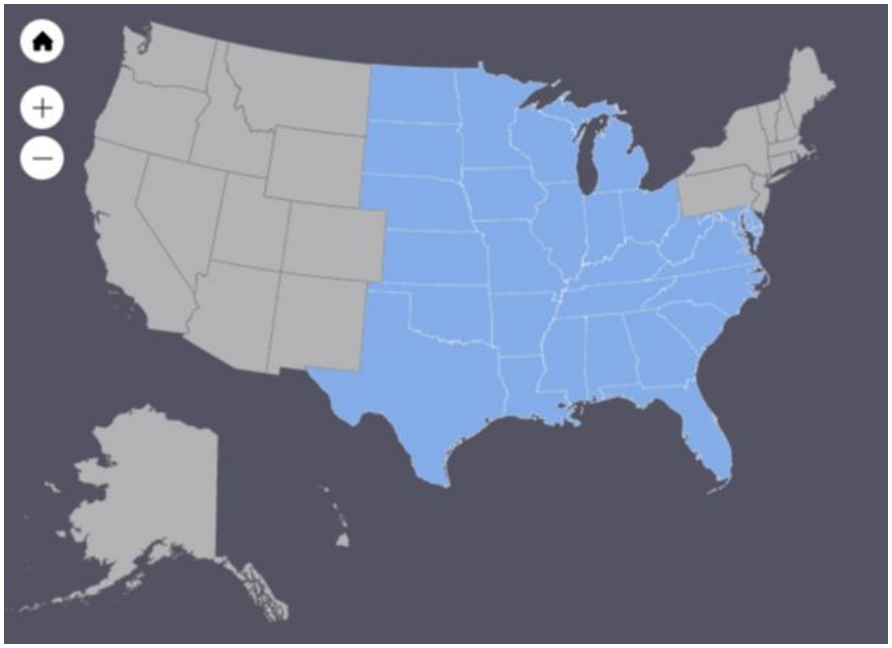
The goal of the logistic regression was to determine if the obesity rate could be used to predict a categorical variable. We began by plotting various discrete variables (determined from thresholds or categories) against obesity rate to see if there were any meaningful relationships that could be explored. One dependent variable we investigated was the state. Plotting the state that the data point was derived from against the obesity rate, we saw that states with lower average obesity rates tended to lie in the west or northeast, whereas states with higher average obesity rates tended to lie in the midwest or south (see below).



We highlighted the top 25 states by obesity rate on the map of the United States, and the geographic trend became much more evident (see below).



To perform a logistic regression on this trend, we constructed a discrete variable, ‘central’, which was 1 for states lying in the Midwest or South census regions of the United States, and 0 otherwise (see below).



For the variable ‘central’, a logistic regression would likely be more useful than a linear regression because ‘central’ only takes on the values 0 and 1, and a logit function more closely

approximates this trend (assuming a linear relationship between the predictor variables and the log-odds of the response variable).

Iteration 1:

For our first model, we decided to see how accurate a regression based on solely average obesity would be. Using the Scikit Learn Logistic Regression package, we generated the following results:

Classification Score (Average Correct Classification Rate): 0.84666879183933696

Coefficients:

1 Avg_obesity [0.40374635762]

We see that an increase in average obesity corresponds to a higher likelihood that the state is central, since the coefficient for average obesity is positive. This is not surprising, since we selected our categorical variable to confirm the trend between obesity and central states.

Overall our classification score was around 85%, but it is important to contextualize this. Since the number of data points given in central states was larger than the number of data points given in noncentral states (2442 > 701), so our null error rate is 0.77696468342. In comparison, our model performed roughly 7 percent better than simply guessing the majority for any given state.

Iteration 2:

For our second model, we added the following variables into our regression model: Perc_Low_Access_To_Store, Avg_groc_store, Avg_super_store, Avg_conv_store, Avg_snap_store, Avg_wic_store, Avg_fast_food, Avg_full_serv, MEDHHINC15, and

Avg_diabetes. This model accounted for all feature variables that were not related to age or race. The results for this model are listed below:

Classification Score: 0.86350037397157819

Coefficients:

1	Avg_obesity	[0.281554356561]
2	Perc_Low_Access_To_Store	[-0.00286076718001]
3	Avg_groc_store	[-0.0186417242319]
4	Avg_super_store	[-0.0366337337684]
5	Avg_conv_store	[0.0209563671292]
6	Avg_snap_store	[-0.00865681874418]
7	Avg_wic_store	[0.0113839089534]
8	Avg_fast_food	[0.0060467962961]
9	Avg_full_serv	[-0.00397226274554]
10	MEDHHINC15	[2.24726480164e-05]
11	Avg_diabetes	[0.523155729792]

Adding in the above variables, we see that most fail to contribute heavily to the model (coefficient < 0.05), which implies that many features, such as the number of grocery stores, super stores, convenience stores, or even the number of fast food restaurants, are roughly equally distributed between central and noncentral states.

Average diabetes, however, contributes to the model, with a higher coefficient than that of average obesity. This suggests that the difference in diabetes rates between central and noncentral states is more significant than the differences in obesity rates. Overall, our second model performed roughly 2 percent better than our initial model, and 9 percent better than the null error rate.

Final Iteration:

In our final iteration, we elected to implement all the given variables in a logistic regression to maximize the classification accuracy. We predicted that adding race and age features would be significant, as the race and age distributions of central and noncentral states are significantly different (shown conveniently by grouping our data frame by 'central' and averaging). The results of this model are given below:

Classification Score: 0.88668661181750186

Coefficients:

1	PCT_65OLDER10	[0.0813561803091]
2	PCT_18YOUNGER10	[0.0630600362731]
3	PCT_NHWHITE10	[0.303632595505]
4	PCT_NHBLACK10	[0.544066477568]
5	PCT_HISP10	[0.306498184588]
6	PCT_NHASIAN10	[0.283725830632]
7	PCT_NHNA10	[0.25467727325]
8	PCT_NHPI10	[-1.88025257092]
9	Avg_obesity	[0.369949651276]
10	Perc_Low_Access_To_Store	[0.000236576152098]
11	Avg_groc_store	[-0.0236568472965]
12	Avg_super_store	[0.0370586627637]
13	Avg_conv_store	[0.0168472839735]
14	Avg_snap_store	[-0.0134470472021]
15	Avg_wic_store	[0.0198546852039]

16 Avg_fast_food [0.00729455405375]

17 Avg_full_serv [-0.00253818413679]

18 MEDHHINC15[1.78767188897e-05]

19 Avg_diabetes [0.283466630339]

Incorporating more feature variables increased the classification accuracy, as expected. Here, we notice that the coefficients for race are all positive (except for Pacific Islander) and relatively large. This suggests that race may play a large role in determining whether or not the given state is central. In contrast, the coefficients for the two given age ranges are positive and smaller than those for race, suggesting that the distribution of ages is marginally wider among central states.

Our final model was able to predict the state of any given data point using the nineteen feature variables listed above with roughly 88.6 percent accuracy, outperforming the null error rate by about 11 percent.

Analysis and Discussion:

For a logistic regression, a high classification score generally indicates that the feature variable set regressed on is a better predictor of the response. Since we began this process by noticing a geographic distribution of states with high obesity, we expected obesity to be one of the strongest predictors of whether a given data point was from a central or noncentral state. Indeed, regressing only on obesity gave a classification score of 84.7 percent.

However, when we began adding other features in our second model, we noticed that average diabetes rate had a higher coefficient in our regression model than average obesity, suggesting that it might be a better predictor of the ‘central’ response. Regressing on the full set

of feature variables, we observed that race was a more significant factor in determining whether a given state was central, as shown by the higher coefficients.

There are several steps we could take to increase the validity of our models. First, to ensure our regression models are not overfitted and are truly accurate, we could split our data into a training set and a testing set. With independent training and testing sets, we would likely benefit from regularization, and we could investigate various methods of regularization to determine the best approach.

Regarding the results of our study, it would be interesting to examine the states classified as ‘central’ and see if there are any underlying causes for general increases in obesity and diabetes (e.g. legislation in these states, cultural preferences, etc.). All of these causes could be analyzed further in future studies to determine if there are clear factors that could potentially be remedied to decrease obesity and diabetes.