Abhinav Pottabathula, Drew Kaul, Kanak Garg, Karim Abedrabbo

Data Science Decal

Project 1

Health vs. Socioeconomic Status in the US:

*Project 1: Food Environment Atlas*

*Introduction:*

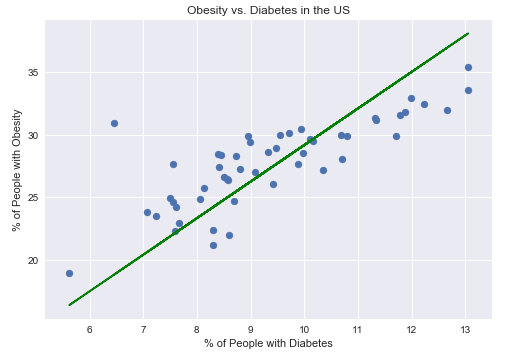
Using the Food Environment Atlas, we extracted data pertaining to how local food choices affects diets and lifestyles in communities in the US. As the Food Environment Atlas was easy to understand and was more in line with our interests, we decided to make it the subject of our project. On top of that, the data was very clean as it contained the least null values and contained many different types of variables, allowing us to pick and choose what we wanted to test for statistical correlation.

*Data:*

We decided to make our main focus in the project to compare the effects of Health, Socioeconomic Status, Prices, Supplemental Data by State/County (Population), Food Insecurity, and Access. First we had to break up the .xls file in which the data originally comes in into multiple CSV’s for the data to be loaded. We took a look at how our data was originally formatted with methods like “.describe(), .head(), .columns, etc,” and we discovered there were very few missing values, but our data was inefficiently formatted for our purposes. Extracting these specific data points, we cleaned information by removing null values and forming arrays for the new data sets. The data was primarily broken down by county in each state, of which there were many counties of data but it was hard to see the overall trend of the nation. With the help of a few helper functions that broke down the unique states of a CSV and reported the amount of counties included in those states, we made a function to average the column values of all the counties according to the state. In this way we effectively made a 51 element state summary for any given CSV. After this, we could more accurately depict our data, however we needed to read the documentation included with the original data to understand what columns correspond to what data.

*Graphing*

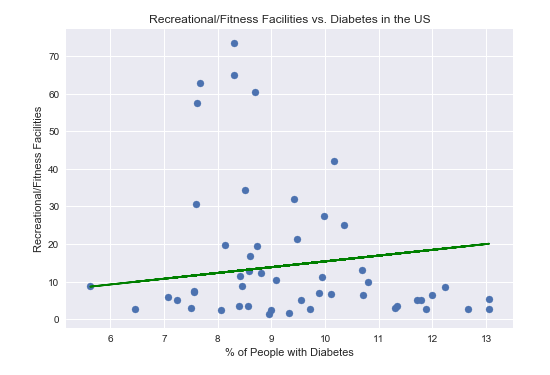
We used state socioeconomic data and state health data for linear regression because it was quantitative and seemed to be relatively closely correlated. Thus we formed the following graphs - Obesity vs. Diabetes, Recreational/Fitness Facilities vs. Diabetes, Overall Poverty vs. Child Poverty Rate, Minority Race Distribution by %, Poverty Rate vs. Median Household Income, Poverty Rate vs. Percent Not in the Labor Force, Poverty Rate vs. Percent Diabetes, and Poverty Rate vs. Percent Obese.

*Conclusions*

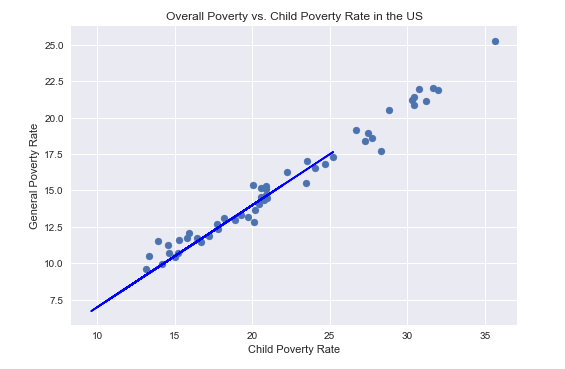
1. Obesity vs. Diabetes

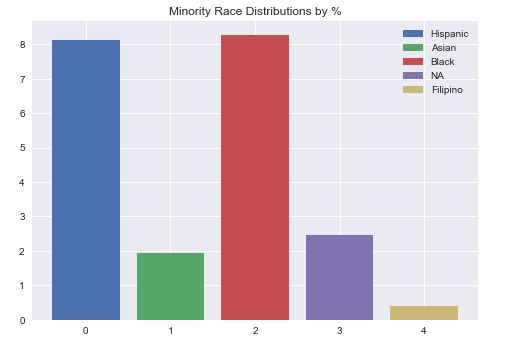
The first graph that we formed was a linear regression comparing Obesity and Diabetes in the US. We found that these two data sets to have a positive and fairly linear correlation where a place with low percent of people with diabetes tends to have a lower percent of obesity and thus a location with higher prominence of diabetes tends to have higher percent of obesity.

B. Recreational/Fitness Facilities vs. Diabetes

The graph below compares the relation between the prominence of diabetes in a certain area against the presence of Fitness Facilities in states around the US. We also applied a linear regression to these data points however we see a data set with a lower correlation and a less strong, while still positive, linear correlation. The data in fact shows no support for the idea that more recreational facilities would lead to a lesser presence of diabetes in our community, which we believed would be the case.

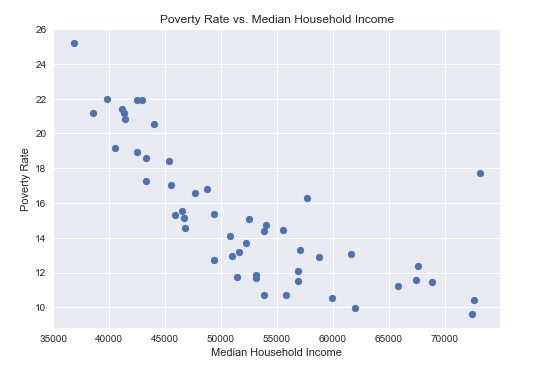
C. Overall Poverty vs. Child Poverty Rate

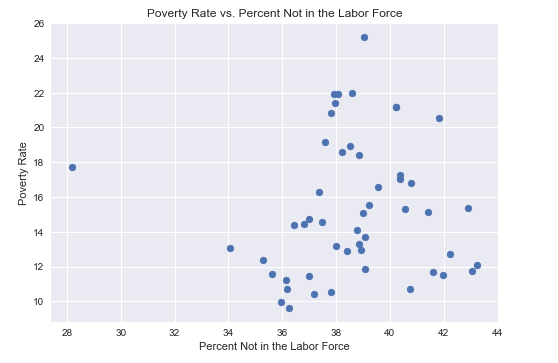
The next graph that we created is another linear regression comparing the relationship between overall poverty in an area to the child poverty rate. As we expected, these data sets followed a very strong linear regression, with a direct relation as when the child poverty rate in any area increases the overall poverty rate tended to increase as well.

D. Minority Race Distribution by %

The next sets of data that we compared was the distribution of Race in respect to the community from where the data was from. This helped us understand that the largest minority groups that our data was delegated from was from Hispanics and African Americans across the board.

E. Poverty Rate vs. Median Household Income

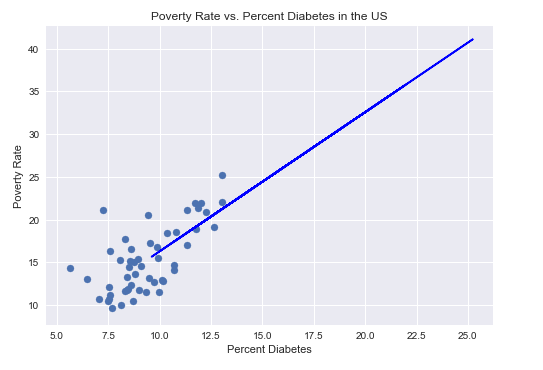
This graph is a scatter plot of data comparing the Poverty rate in communities based on how high the Median Household Income. Though the data seems to have a strong, negative correlation hinting that as the median household income decreases, the poverty rate increases and vice versa, we decided not to include a linear model as the residual graph of this data had a pattern, demonstrating that this potential correlation is not linear.



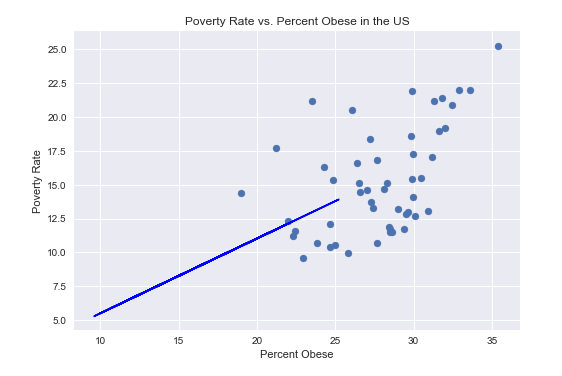
F. Poverty Rate vs. Percent Not in the Labor Force

The graph of poverty rate against the percent not in labor force data, we were able to see very little to no correlation as the data represented a set of data with random scatter.

G. Poverty Rate vs. Percent Diabetes

The next graph is a linear regression comparing data of poverty rate in relation to percent of people with diabetes in a community. The graph seems to have a positive, somewhat strong correlation which seems to demonstrate that with an increase in diabetes, there seems to be an increase in poverty and vice versa.

H. Poverty Rate vs. Percent Obese

The final graph is a linear regression which plots the percent of obese people in a community against the poverty rate. Based on the graph, there seems to be a positive and somewhat weak correlation discerning that with the increase in percent of obese people in an area, the poverty rate seems to have some positive correlation. As percent obese increase, so does the poverty rate, and vice versa.

*Logistic Regression*

We attempted to create a graph using logistic regression; however, it wasn’t possible. Logistic regression requires categorical data sets which we can compare to create a sigmoid curve and thus analyze how probable one category can occur based on other data. However, with this data set, and by the time we finished up linear regression, we could not find any source of categorical data that would fit the requirements for logistic regression as all our data was found in percents or percent changes. In the case that we were provided data such as health data that was more qualitative, we would be able to conduct logistic regression. For example, based on someone’s weight/level of obesity, race, age, gender, and other factors, would that person be diabetic (creating a threshold so if the risk of being diabetic was over x% they would be considered diabetic)? We could model the data and determine based on the individual data sets if one is likely or not to be diabetic based on these traits. A specific example would be to analyze obesity based on the number of meals a person eats. What we could do is plot whether certain men are obese or not against the number of meals and thus understand how many meals can contribute to what probability of getting obese. Fitting this data to a sigmoid curve would thus demonstrate this. However, we did not have any data that matches these requirements, and thus we were not able to create a graph for logistic regression.