Predictive Analysis for Diabetes ( for the US Population)

# Project Overview:

The project seeks to analyze factors that lead to occurrence of diabetes and makes the population more susceptible to this life style disease. We also seek to find why the more of US population falls in the net of diabetes with every passing year. This disease has been increasing its spread with the year although the corresponding rise in the population and the demographics has been almost negligible.

# Sources:

The data was taken from the USDA website considering that the data is credible as it comes from a government department and is free sourced and published online. The link is as given below:

<https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/#Current%20Version>

# Objectives:

* Try to figure out factors that can be a leading cause of diabetes and their affect.
* The factors include obesity, access to fresh / better quality food, nearness to SNAP authorized stores, Restaurants (fast food and full service), farmer markets (along with those accepting SNAP)
* How does each of the factors affect in lowering of increasing of occurrence.
* What steps can be taken to improve the situation

Note: The Supplemental Nutrition Assistance **Program** (**SNAP**), formerly and commonly known as the Food Stamp **Program**, provides food-purchasing assistance for low- and no-income people living in the United States.

# Requirements/Task(s):

* Download the data from a credible source.
* The workbook downloaded contained multiple worksheets. All of them need to cleaned, relevant features selected from each of the csv’s, dropping data that was from before 2012.
* Since research is done every 4-5 years, data for stores, restaurants etc. (from 2014 – 2016) was considered credible and percentage change was taken for most of the features.

# Notes:

Obesity came out as a major factor from the data.

# Outline the steps/plan for your project:

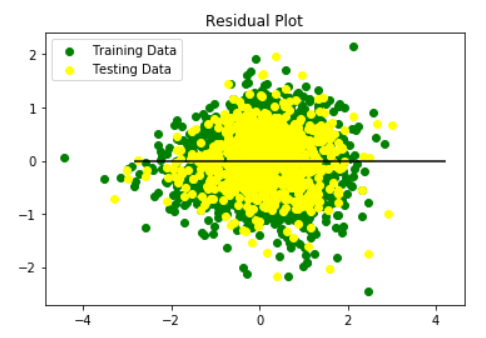
* Cleanup the data using python, pandas, excel.
* Fill up / remove NAN values, select relevant columns, rename and compute where ever required.
* Save to CSV files.
* Merge all resulting csv’s into single csv file so that it can be used for SKLearn and Tableau analysis.
* Use multiple linear regression analysis from SKlearn to analyze the features
* Use superwised learning with test, train, split functions.
* Scale, fit and score the model.
* get the weights

# Summarize what you learned:

The training and testing score showed that the occurrence can be predicted as correct 70% of the time if the features on which it is being predicted are the same as those used to create the model.

Training Score: 0.7023462457014373

Testing Score: 0.7131523555249148



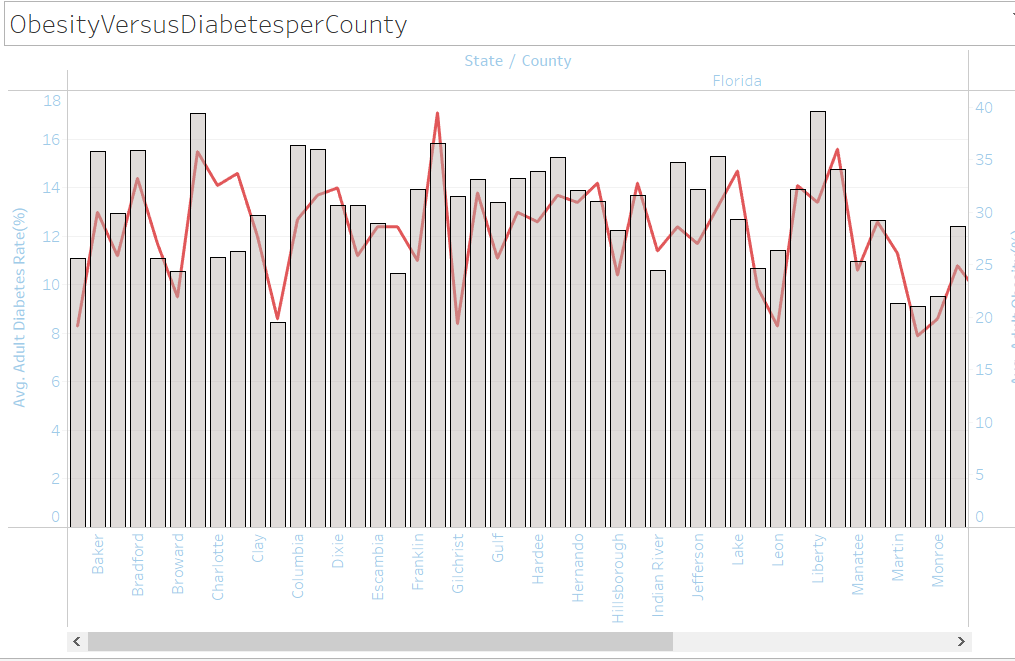
The MSE and the R-square values came as follows

Mean Squared Error (MSE): 0.28343240317564106

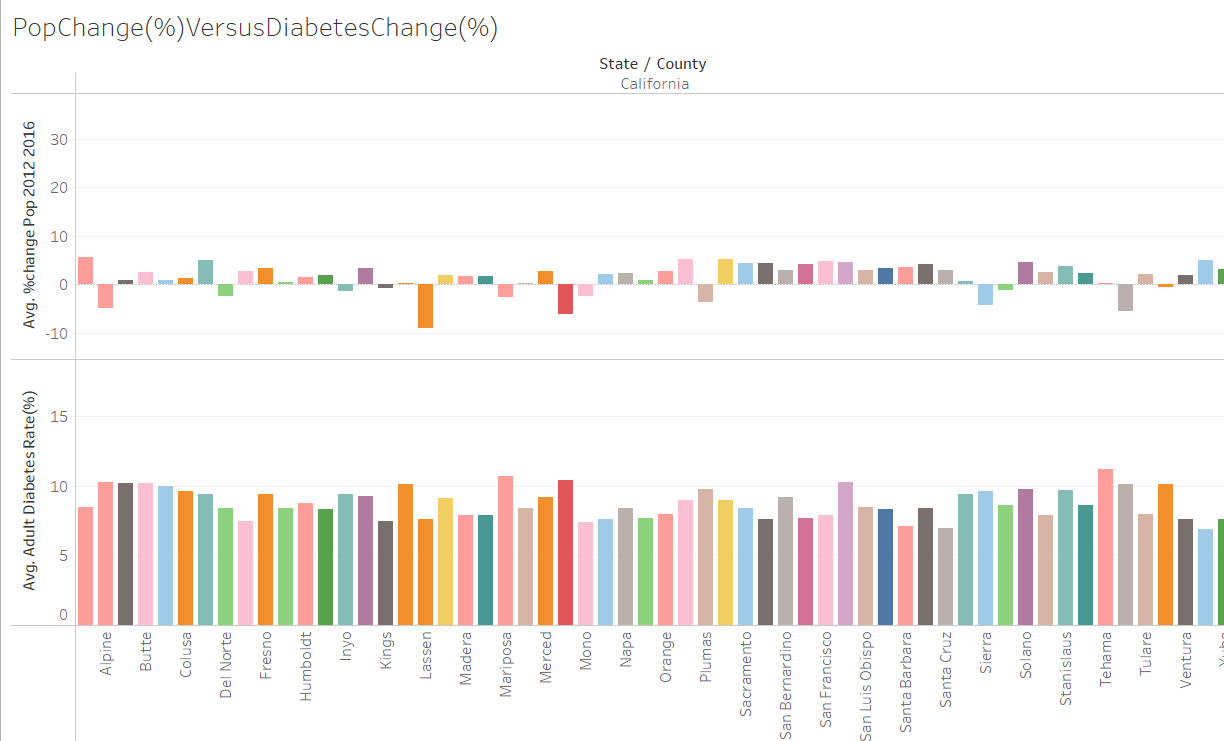
R-squared (R2): 0.7131523555249148

Tableau Analysis:

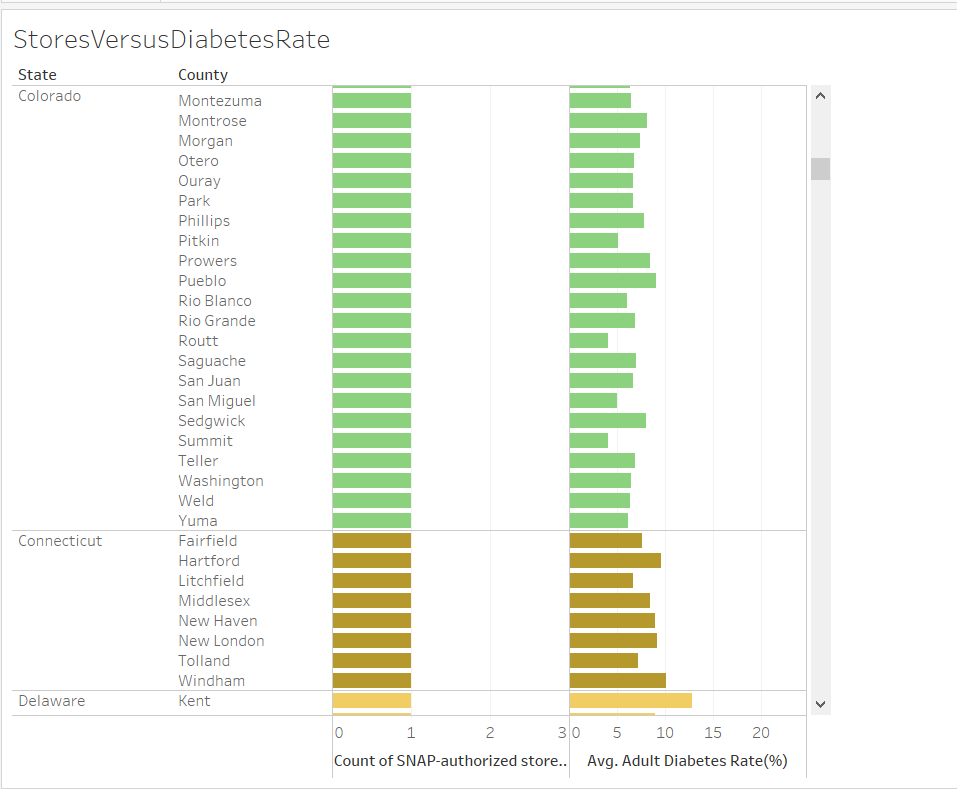
Link to Tableau Public : <https://public.tableau.com/profile/madhulika.gupta#!/vizhome/diabetesPredictionAnalysis/PopChangeVersusDiabetesChange?publish=yes>



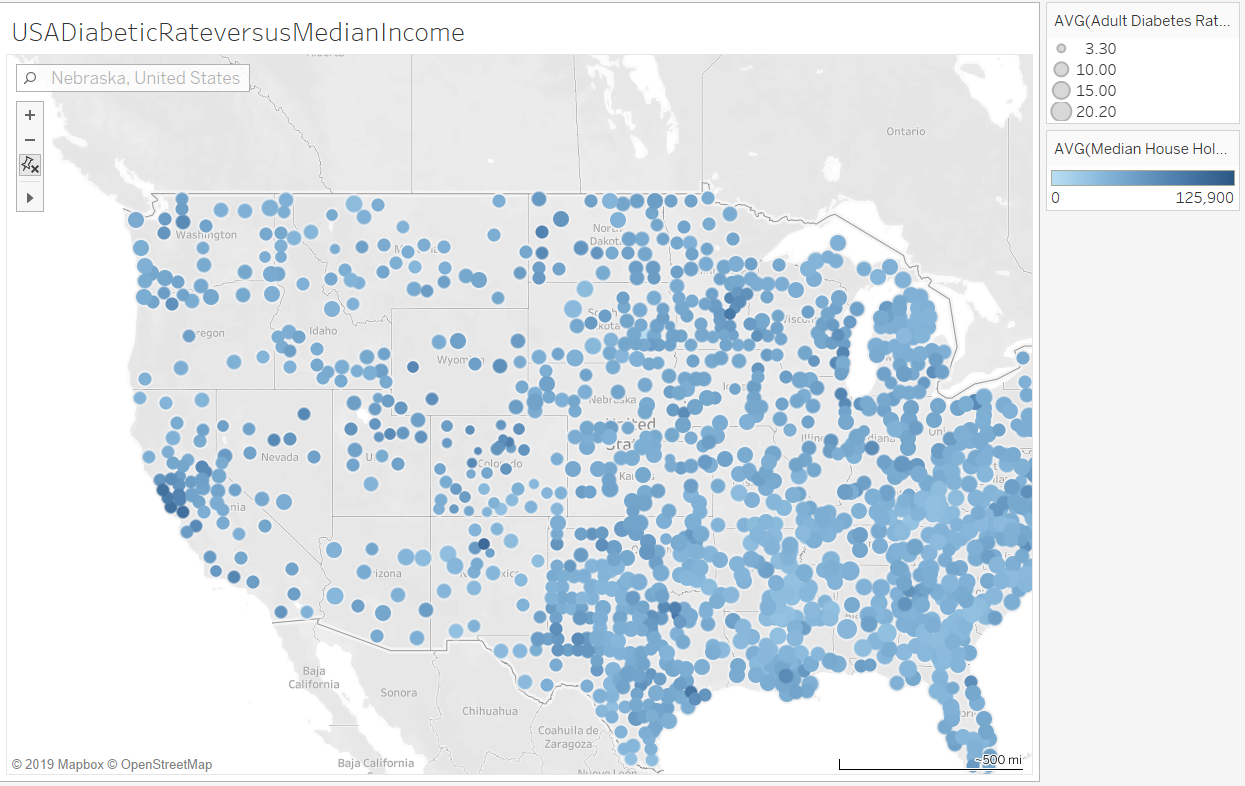
High rates of obesity result in high occurrence of diabetic rates



Diabetes is rising at huge rates even as change in population remains negligible in most counties (some places show negative growth and increase in diabetic rates)



The no. of restaurants, SNAP authorized centers, market accepting redemption stamps remain almost constant.



This one show Diabetic rates versus median household income. The tooltips would show that diabetic rate are lower where ever the median house hold income is higher.

Since obesity is a major factor, if we can improve access to more nutritional food (whether thorough, SNAP accepting centers and markets, more access to gyms motivating to be physically active etc. should help to improve the situation)