PROJECT 1

Correlation between Income and Health in United States

ECEN 689: Applied Information Science Practicum

Course Instructor: Prof. Jean-Francios Chamberland

TEAM 4

Harish Chigurupati
Drupad Khublani
Sambandh Bhusan Dhal
Kishan Shah

***Abstract*:** The goal of the project is either to support or refute the hypothesis that "Income level is positively correlated to health". Thee datasets comprising of the diabetes and obesity rates for two years 2008 and 2013 were taken as reference datasets to select the features which we use in this project to find out the relationship between them.

I. INTRODUCTION

The datasets which we use to start with the project are basically divided into 2 parts:

1. Food Atlas dataset provided by the US Department of Agriculture which contains statistics related to food environment indicators such as Adult Diabetes Rates and Adult Obesity Rates for the years 2008 and 2013 along with Recreation and Fitness Facilities for the years 2009 and 2014.
2. The Income Tax Statistics Dataset provided by the United States Federal Government provides an estimate of the income and the tax paid by the individuals in the year 2008 and 2013. The dataset contains 144 columns.

Using the datasets given, we try to analyze and bring out a conclusion whether health plays a role in determining one's income in the United States of America.

According to [1], the general notion is that poor people have worse health because they have insufficient money and people who are earning more would have better health conditions but there has been a long debate as to the merits of describing poverty in absolute or relative terms. As stated in the paper[3], it is worth distinguishing two aspects of low income: "poor material conditions" and "lack of social participation". We use this paper as a primary concept to do exploratory analysis on our data.

According to [2], this paper too reiterates the first observation. The only difference in the two papers are the predictors which they have used for analysis of the data. Here, they go on to use:

1. Materialistic arguments
2. Psychological mechanisms
3. Behavioral factors
4. Influence on education and employment opportunities.

Based on these journal papers mentioned above and the datasets which we had considered, we tried to investigate and find the correlation between health and income among the people belonging to different socio-economic status in the United States of America

II. DATA PRE- PROCESSING

- Food Atlas Dataset:

We considered 6 columns from this dataset *i.e.* the obesity and diabetes parameters for the years 2008 and 2013 and recreation and fitness facilities for 2009 and 2014 which were categorized by their corresponding FIPS county code.

- IRS Dataset:

Here, we consider 1 column as the main defining characteristic of the dataset namely Adjusted Gross Income. It has been given corresponding to zip code in U.S.

- US Zip codes to County State to FIPS Crosswalk

We use this dataset [4] as mean to merge both the above two datasets to get our final dataset.

III. LINEAR REGRESSION

To get some quantitative measure for the correlation between health and income we use linear regression algorithm. We use Obesity and diabetes from 2013 and Recreation facilities from 2014 as our predictor and adjusted gross income as our response. From the slopes we can clearly say that obesity and diabetes are negatively related to income while recreation facilities is positively correlated to the income.

Coefficients for the model used are as follow:

% Diabetes 2013: (-18098.74)

% Obesity 2013: (-8503.67)

Recreation Facilities 2014: (7102.75)

We also tried to show the correlation matrix below.

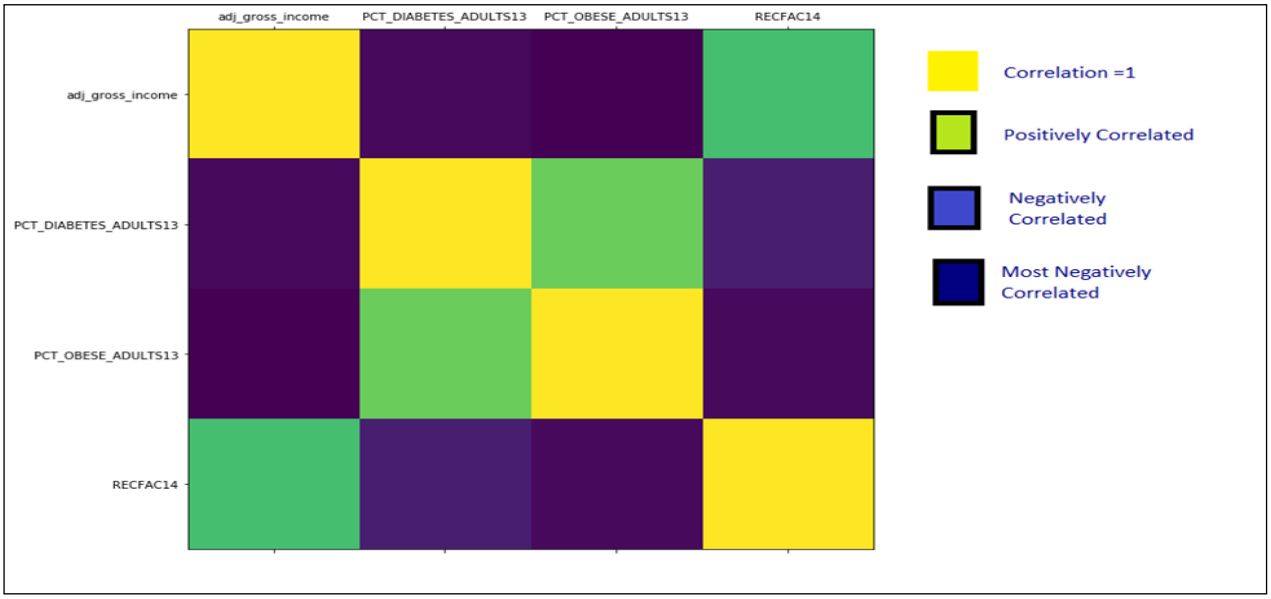
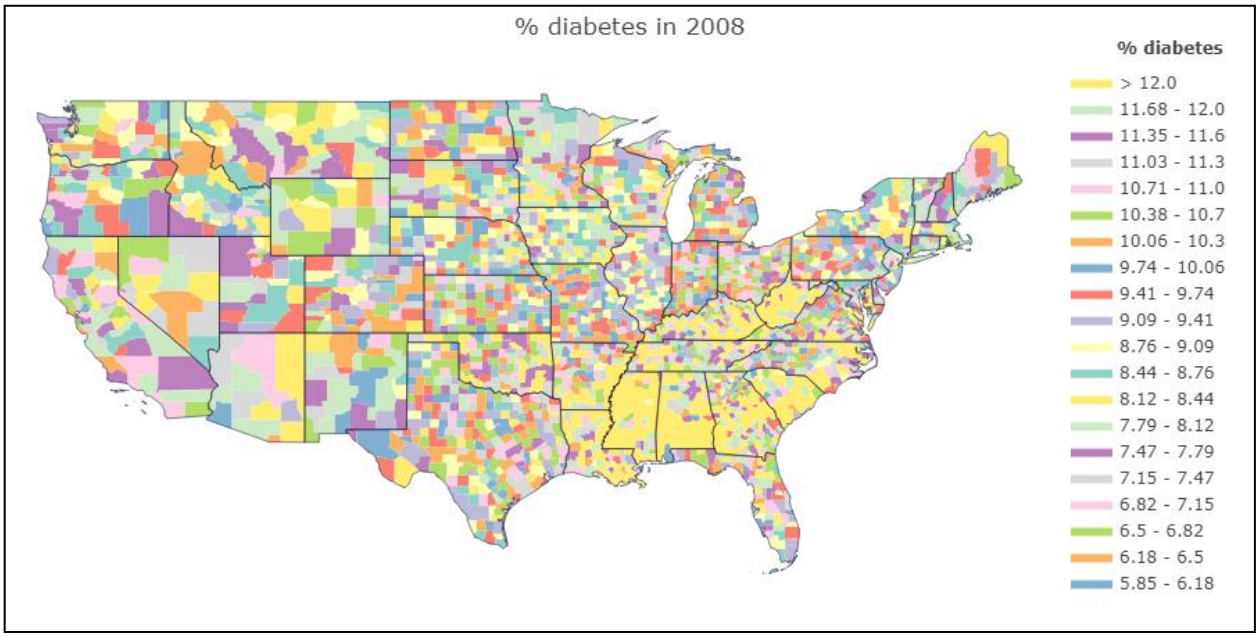
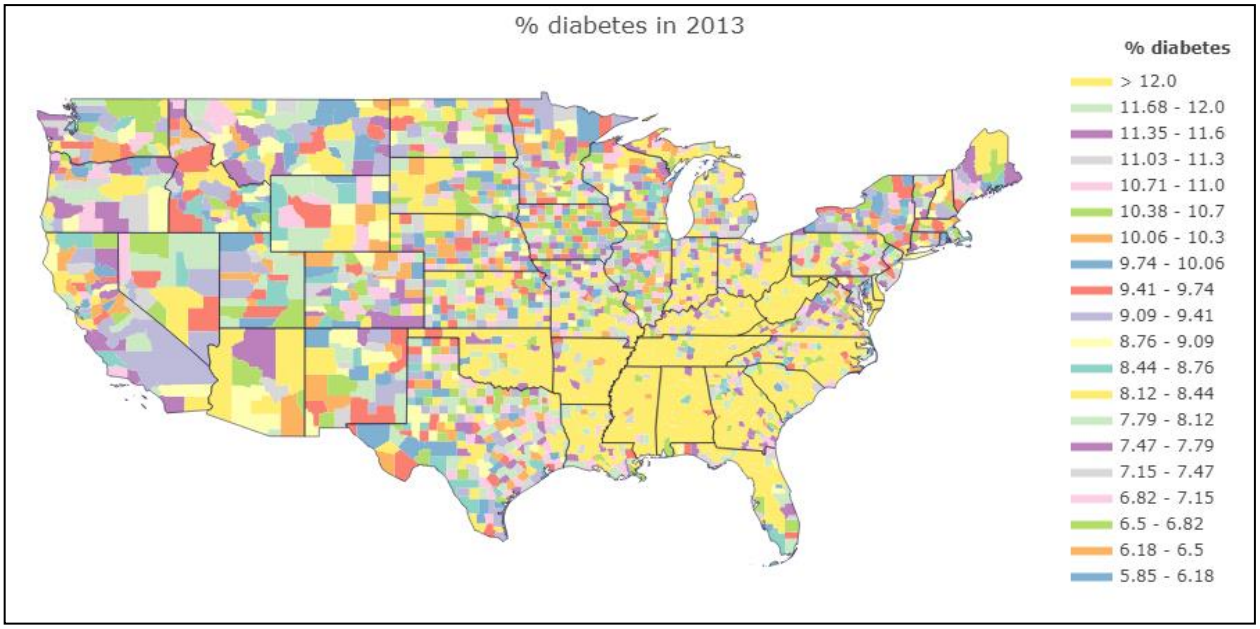


Figure 1 - Correlation matrix of linear regression model

IV. EXPLORATORY DATA ANALYSIS

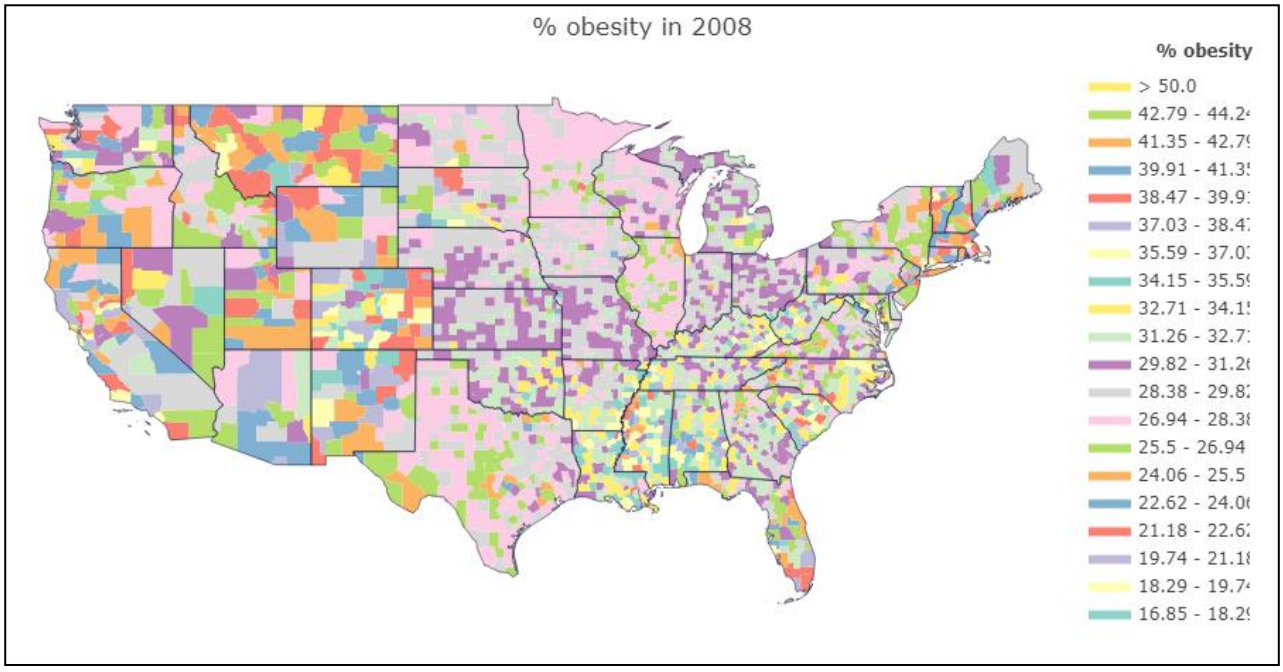
- Analysis of Diabetes in the US population for the years 2008 and 2013:

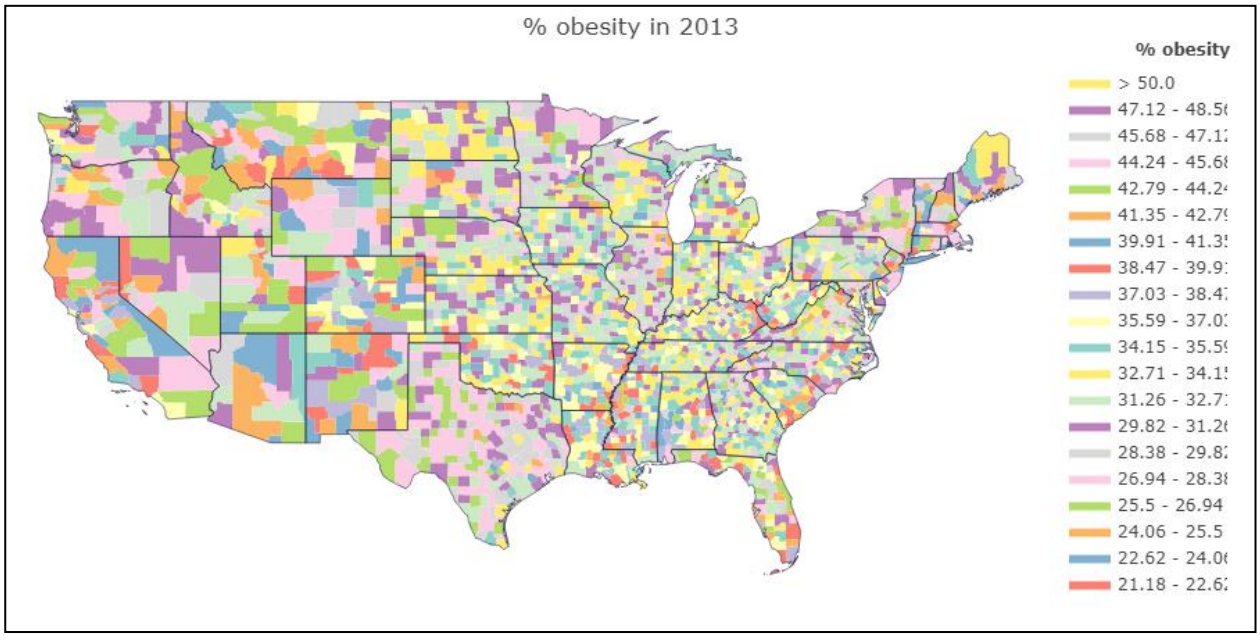




From the year 2008 to 2013, we can conclude that the percent of population having diabetes has considerably increased in the South-Eastern States in United States of America. i.e. the states of Mississippi, Alabama, Georgia, Northern Florida, Tennessee, Kentucky, Arkansas, Louisiana, Oklahoma, West Virginia, North Carolina, South Carolina and Ohio. These states have shown tremendous increase in the percentage of diabetic population in a span of 5 years. The only North-Eastern state which can be clubbed into this group is Maine. overall, almost every state in the USA has shown considerable increase in diabetes.

- Analysis of Obesity in the US population for the years 2008 and 2013:

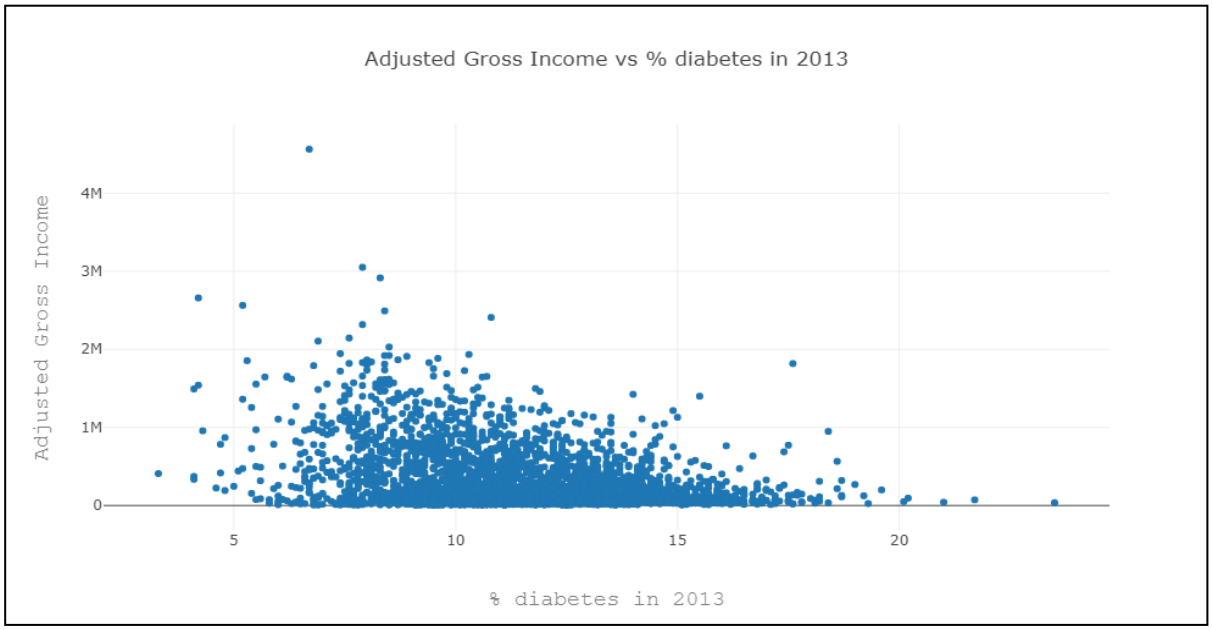




In the year 2008, we can notice that the obesity rates are somewhat constant in the Central states of America but in the year 2013, the percentage of people having obesity is scattered. The states of Arkansas, Louisiana, Oklahoma, Mississippi and Iowa showed the highest percentage of obese people.

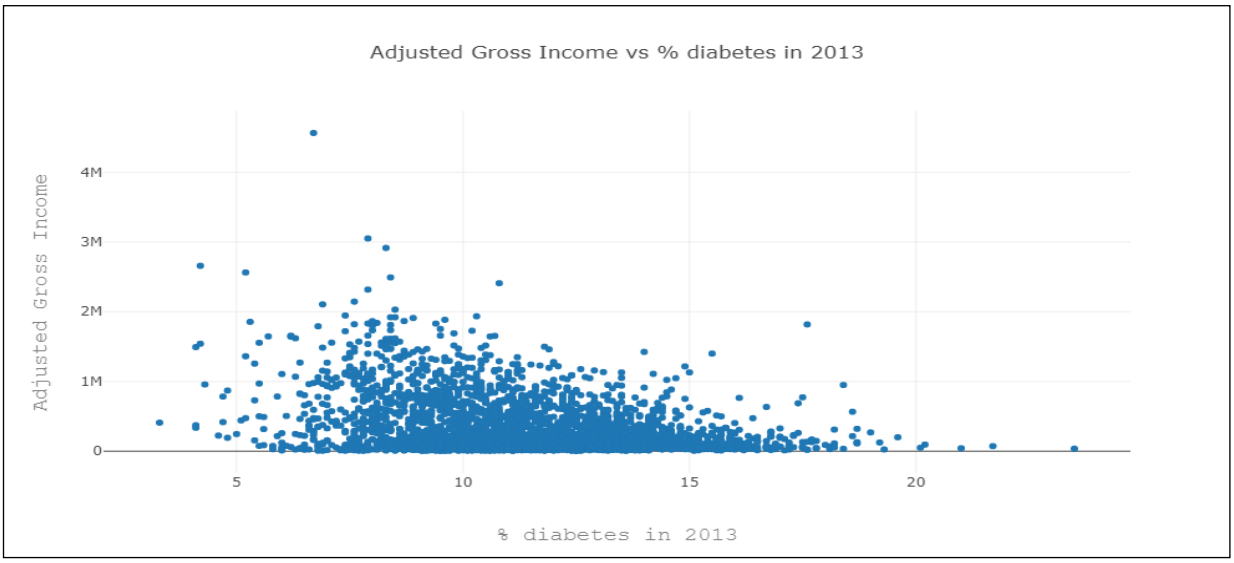
- Adjusted Gross Income Versus Health parameters:

(A). Adjusted Gross Income versus Diabetes in 2013:



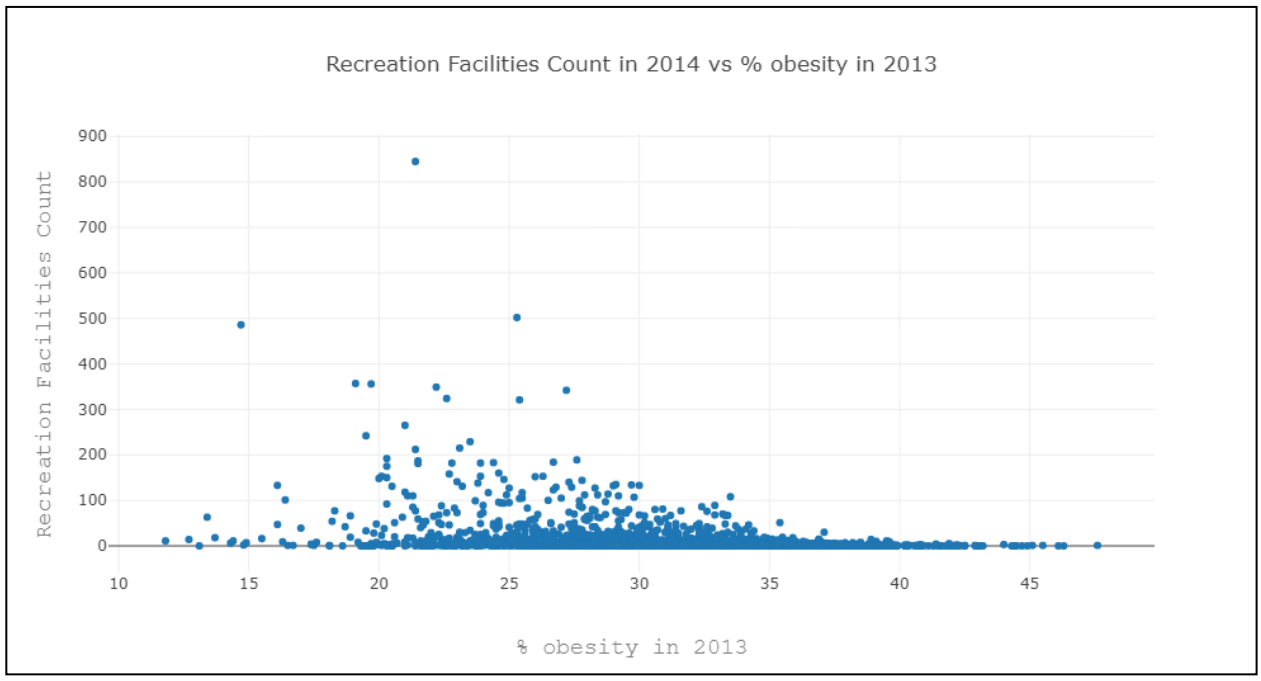
From the graph, we can see that the trend more or less remains the same i.e. the people with the lowest adjusted gross income seem to be the most diabetic people.

(B). Adjusted Gross Income versus Obesity in 2013:



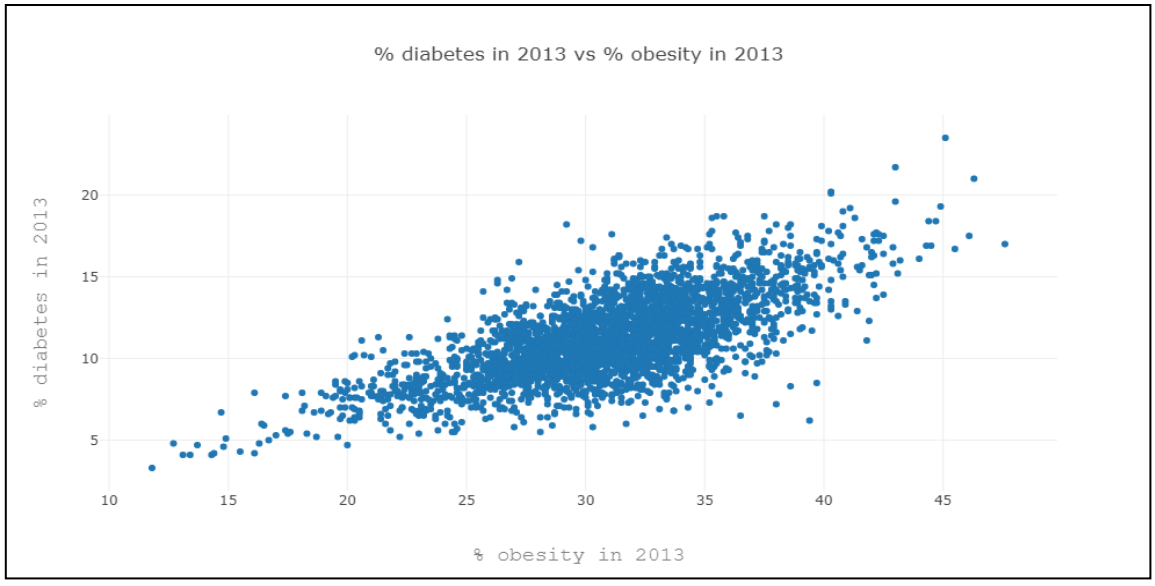
From the graph, we can see that the trend more or less remains the same i.e. the people with the lowest adjusted gross income seem to be the most obese people.

- Recreational Facilities Count versus Obesity and Diabetes



From the above graphs, we see that when the recreational facilities count increases, the obesity rate and diabetes rate decreases. This holds good for both the years 2009 and 2014.

- % Diabetes v/s % Obesity in 2013:



From the graph, we can conclude that the percentage of diabetic patients go onto increase with the increase in the rate of obesity among the population. The graph shows a linear relationship.

V. CONCLUSION

In the end, based on our exploratory data analysis and results from machine learning algorithms we can conclude that income is positively correlated to good health of an individual.
The health and income datasets estimate of the people belonging to different strata of the society was taken into consideration for the years 2008 and 2013. We analyzed the data to take certain predictors into consideration and see how the different parameters influence the health and income of individuals.
Features which reflect health of a person like percent diabetes and obesity have increased from the year 2008 to 2013 mostly in the Southern states of United States of America. As diabetes and obesity rate have a negative correlation on income, it can be stated that income is positively correlated to the good health of an individual.

VI. REFERENCES

[1]. Michael Marmot, the Influence of Income on Health: Views of an Epidemiologist.

[2]. Micaela Benzeval, Lyndal Bond, Mhairi Campbell, Mathew Egan, Theo Lorenc, Mark Petticrew, Frank Popham, How does Money Influence Health?

[3]. Adam Wagstaff, Eddy van Doorslaer, Income Inequality and Health: What does the Literature Tell Us?

[4]. https://www.kaggle.com/danofer/zipcodes-county-fips-crosswalk/version/1#_=_