Exploratory Data Analysis of Diabetes in the United States of America

Markanday Ravi, Mahsa Mirzargar, Burton Rosenberg
Department of Computer Science, University of Miami

Introduction

Diabetes is a highly preventable, debilitating disease that is expensive to address, with an economic cost of about $245 billion as of 2012. It is growing at a rapid rate, from 23.6 million people, or 7.8% of the population in 2007, to about 30.3 million people, or 9.4% of the US population as of 2015.

It is beneficial to develop a higher level perspective as to what influences the development of this disease, in order to make it easier to identify which details are important. This can help with developing and testing theories for the various ways in which the disease develops for different subsets of the population.

The objective of this research is to understand how factors outside of genetics, with a focus on income, are related to the development and spread of diabetes. The aspects that may be related include income, geographical location, ethnicity and gender. This analysis can provide an insight into the relationships between these factors and the disease. Continuation of this research may provide a greater understanding of the current situation, and how to address it. The current conclusion of this research is that it is statistically significant that there exists a relationship between income and diabetes.
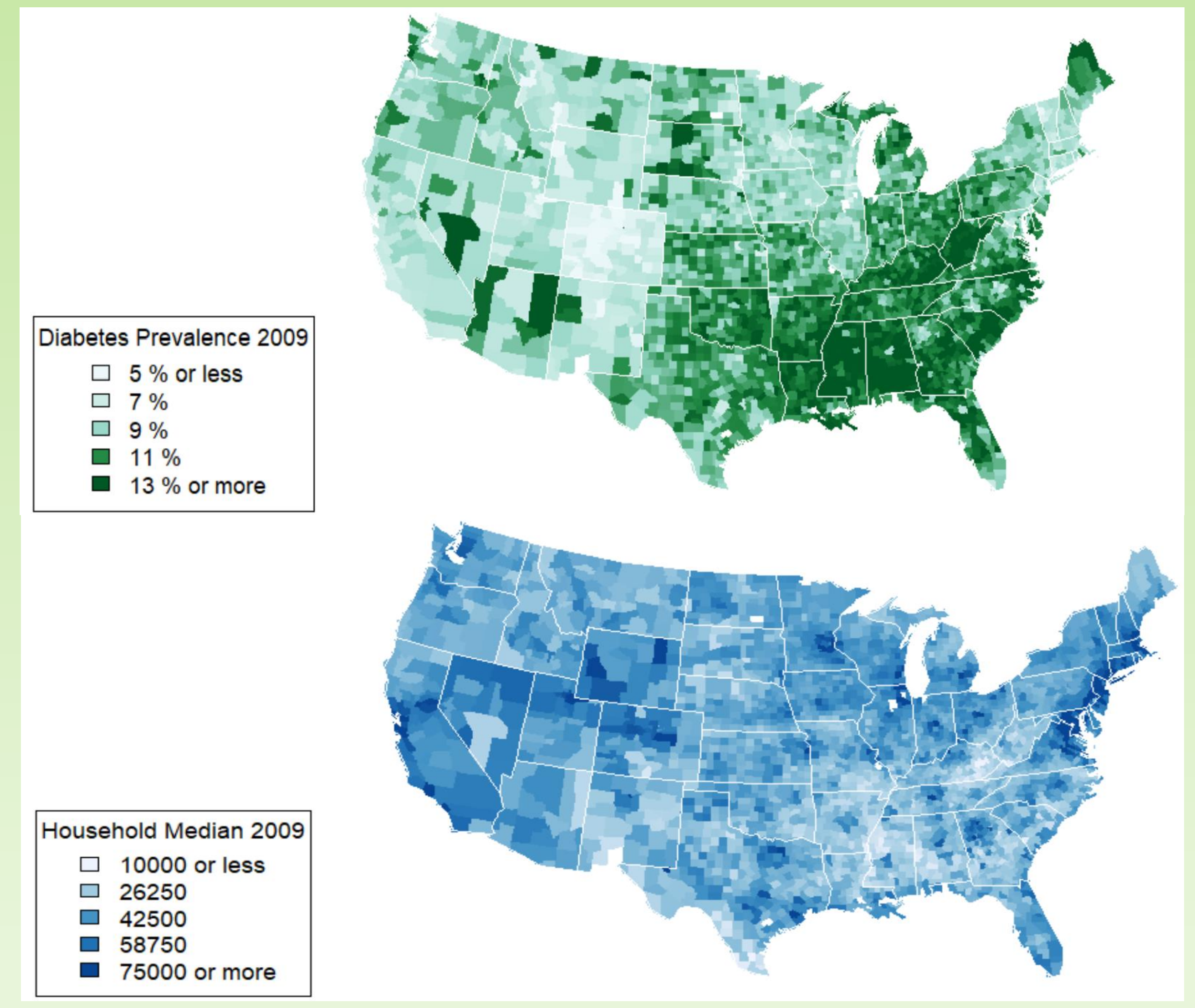
Methods

- Income data was collected from the Census Bureau website, for the years 2009-2015
- Diabetes data was collected from the CDC website, for the years 2004-2013
- Data were converted to CSV files before being used as input
- Datasets were processed, visualized, modeled, and analyzed using R.
- Modified datasets were mapped using Shiny in R and the Leaflet package

Results

Visualization
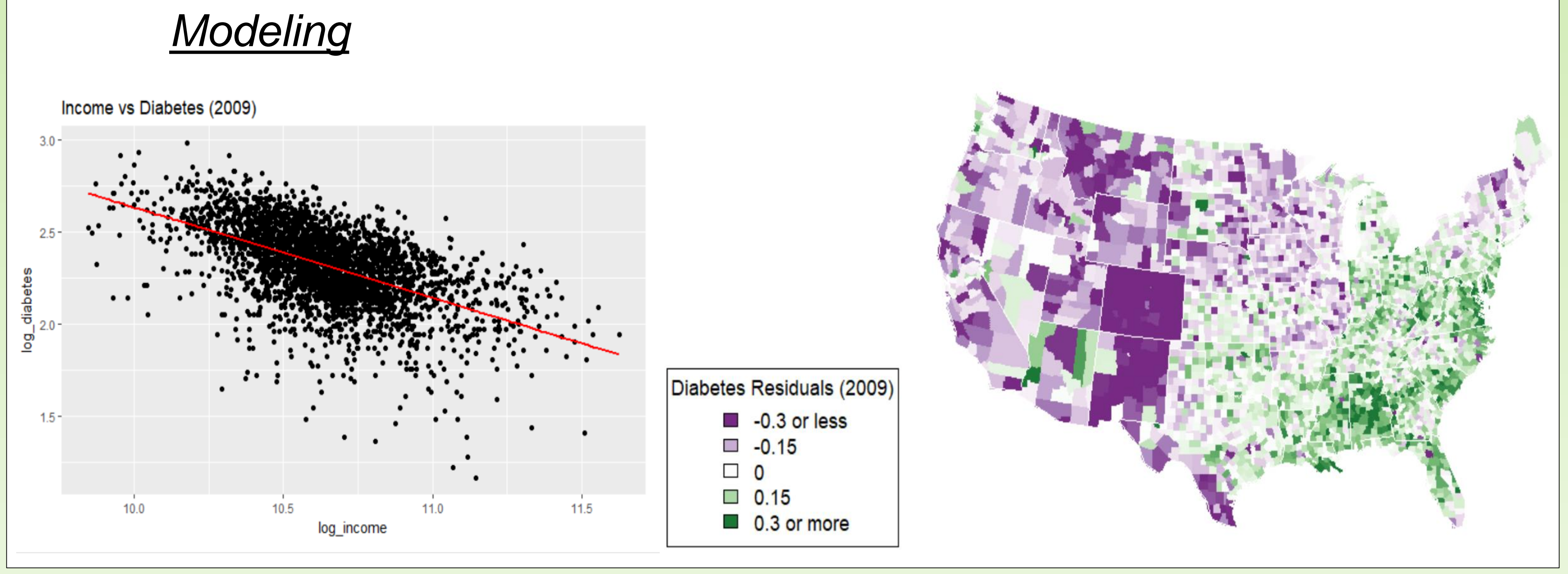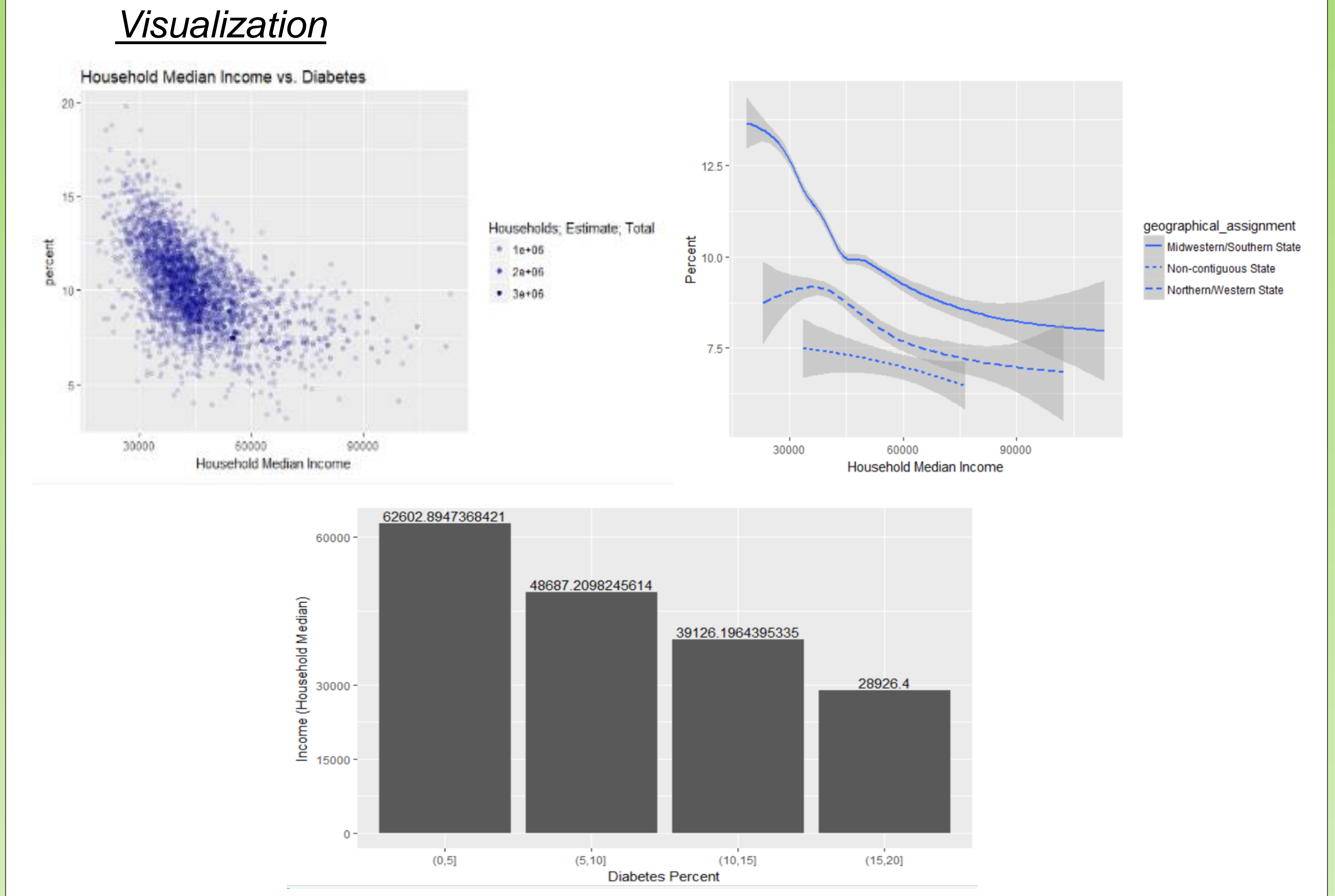
Modeling

Conclusions

- Within R, the function **lm** [lm(formula = log_diabetes ~ log_income] determines the preposition "log_income has no linear effect on log_diabetes", as the null hypothesis
- P value of the model is **2.2 e -16**.
- In this sample of data, it is highly statistically significant that median household income has a linear effect on the diabetes prevalence percent of a county.

Future Research

- Explore whether geography influences the strength of the correlation between income and diabetes
- Explore the strength of correlation between income and diabetes between genders
- Analyze whether poorer households in wealthier counties fare better than poor households in poor counties
- Analyze the relationship between income and diabetes throughout multiple years
- Can also check whether the average education level of county affects the correlation between income and diabetes
- Explore whether diabetes incidence, defined as the rate of new cases of diabetes per year, follow a similar relationship between the various factors

Acknowledgements