CS498 Cloud Computing Applications

Project Proposal – U.S. Health Data Analysis

Team Members (Team 30)

Nicolas Maire (nmaire2)– Basel, Switzerland
John Moran (jfmoran2) – Dalian, China
Graham Chester (grahamc2)– London, UK

Table of Contents

1. Introduction and Background

We propose to analyze United States health datasets as released by the Centers for Disease Control and Prevention (CDC), and the Health Inequality Project, using the Hadoop ecosystem. We will use this data to both identify the healthiest and unhealthiest places to live in the US, and to analyze key relationships between health risk factors and outcomes, income and sex by major US City. The World Health Organization has declared reducing health inequities an important goal, and the US has one of the least equitable health systems in the industrialized world (1). We believe the intellectual merit of this study lies in developing a better understanding of these inequalities and impacts on health in the United States today.

2. Datasets and Toolsets

There are a broad variety of datasets from varying sources (of varying quality) available regarding aspects of health in the United States. The datasets we have chosen firstly are sourced from quality organizations, and secondly, enable us to focus on the impacts of location and socio-economic status on individual health.

The 500 Cities Dataset (2) is released annually by the CDC (https://www.cdc.gov) from a collaboration between the CDC, the Robert Johnson foundation and the CDC Foundation. It provides 27 key health risk factors and outcomes (listed in Appendix A), for 500 cities and census tracts within the United States.

The Health Inequality Dataset (3) provided by the Health Inequality Project (https://healthinequality.org), contains life expectancy by US city for men and women by income quartile.

We propose to use the following tools from the Hadoop ecosystem to analyze the dataset:

- Amazon EC2 for Linux instances
- Hortonworks distribution to provide HDFS, Spark and supporting tools
- Spark RDD's and Spark Shell for interactive manipulation and analyses of the data
- Spark MLlib for correlations and other analyses.

3. Proposed Analysis and Extensions

We propose to provide an analysis of the most and least healthy places in the US to live, using a set of weights on the measurements defined in Appendix A in order to calculate a health score. We will attempt to define these weights to align with public perception of the criticality of each of the measurements.

We will then incorporate another data set: Health Inequality (3) comprised of life expectancy, gender and income data by U.S. city. With the data sets joined, we will examine various correlations with the expectation to have solid data confirm intuitive assumptions regarding the relationships between poverty and issues such problem drinking, obesity, and lack of preventative health care.

The purpose of using a cloud ecosystem and Hadoop big data cluster for this project is to allow additional large scale datasets to be added to the analyses without significant technical impact. For example, the analysis could be extended to additional geographic areas outside the 500 cities in the US, beyond the US, or possibly include more detailed salary data by city to further investigate socio-economic impacts on health. We will demonstrate scaling up either by including additional data sets such as these, or by using synthetic data if this is not feasible.

4. References

1) Davis K, Stremikis K, Squires D, Schoen C. Mirror, mirror on the wall: how the performance of the U.S. health care system compares internationally. New York, NY: The Commonwealth Fund; 2014
2) 500 Cities Dataset (2017). Retrieved on 07-Feb-18 from https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2017-relea/6vp6-wxuq
3) Health Inequality (2017). Retrieved on 18-Feb-18 from https://healthinequality.org/dl/health_ineq_all_online_tables.xlsx

Appendix A – 500 Cities Dataset Definition

The 500 Cities Dataset (1) in the 2017 revision contains data from 2014 and 2015, and consists of 810,103 rows with the following fields.

Column Name	Description	Type
Year	Year	Number
StateAbbr	State abbreviation	Plain Text
StateDesc	State name	Plain Text
CityName	City name	Plain Text
GeographicLevel	Identifies either US, City or Census Tract	Plain Text
DataSource	Data source	Plain Text
Category	Topic	Plain Text
UniqueID	At city-level, it is the FIPS code of CityFIPS. For tract-level data, it is a combined ID of CityFIPS and TractFIPS for tracts within the respective city with the exception of Honolulu, which only uses TractFIPS	Plain Text
Measure	Measure full name	Plain Text
Data_Value_Unit	The data value unit, such as "%" for percentage	Plain Text
DataValueTypeID	Identifier for the data value type	Plain Text
Data_Value_Type	The data type, such as age-adjusted prevalence or crude prevalence	Plain Text
Data_Value	Data Value, such as 14.7	Number
Low_Confidence_Limit	Low confidence limit	Number
High_Confidence_Limit	High confidence limit	Number
Data_Value_Footnote_Symbol	Footnote symbol	Plain Text
Data_Value_Footnote	Footnote text	Plain Text
PopulationCount	Population count from census 2010	Number
GeoLocation	Latitude, longitude of city or census tract centroid	Location
CategoryID	Identifier for Topic/Category	Plain Text
MeasureId	Measure identifier	Plain Text
CityFIPS	FIPS code	Plain Text
TractFIPS	FIPS code	Plain Text
Short_Question_Text	Measure short name	Plain Text

Health Risk Factors and Health Outcomes are percentages of the following within the population:

Health Risk Factors
Binge drinking among adults aged >=18 Years
Visits to doctor for routine checkup within the past Year among adults aged >=18 Years
Cholesterol screening among adults aged >=18 Years
Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50–75 Years
Older adult men aged >=65 Years who are up to date on a core set of clinical preventive services: Flu shot past Year, PPV shot ever, Colorectal cancer screening
Older adult women aged >=65 Years who are up to date on a core set of clinical preventive services: Flu shot past Year, PPV shot ever, Colorectal cancer screening, and Mammogram past 2 Years
Current smoking among adults aged >=18 Years
Visits to dentist or dental clinic among adults aged >=18 Years

No leisure-time physical activity among adults aged >=18 Years
Mammography use among women aged 50–74 Years
Papanicolaou smear use among adult women aged 21–65 Years
Sleeping less than 7 hours among adults aged >=18 Years
Health Outcomes
Arthritis among adults aged >=18 Years
High blood pressure among adults aged >=18 Years
Taking medicine for high blood pressure control among adults aged >=18 Years with high blood pressure
Cancer (excluding skin cancer) among adults aged >=18 Years
Current asthma among adults aged >=18 Years
Coronary heart disease among adults aged >=18 Years
Chronic obstructive pulmonary disease among adults aged >=18 Years
Physical health not good for >=14 days among adults aged >=18 Years
Diagnosed diabetes among adults aged >=18 Years
High cholesterol among adults aged >=18 Years who have been screened in the past 5 Years
Chronic kidney disease among adults aged >=18 Years
Mental health not good for >=14 days among adults aged >=18 Years
Obesity among adults aged >=18 Years
Stroke among adults aged >=18 Years
All teeth lost among adults aged >=65 Years

Appendix B – Health Inequality Dataset Definition

From the available Health Inequality datasets, we utilize Online Data Table 6 which contains "Commuting Zone" level life expectancy (at age 40) estimates for men and women by income quartile. Key fields are shown below.

Column Name	Description	Type
cz	Commuting Zone ID	Number
czname	Commuting Zone Name	Number
pop2000	Commuting Zone Population in 2000	Number
fips	State FIPS	Plain Text
statename	State Name	Plain Text
stateabbrv	State Abbreviation	Plain Text
le_raceadj_q1_F	Race-Adjusted, Q1, Female Life Expectancy	Number
le_agg_q1_F	Unadjusted, Q1, Female Life Expectancy	Number
le_raceadj_q2_F	Race-Adjusted, Q2, Female Life Expectancy	Number
le_agg_q2_F	Unadjusted, Q2, Female Life Expectancy	Number
le_raceadj_q3_F	Race-Adjusted, Q3, Female Life Expectancy	Number
le_agg_q3_F	Unadjusted, Q3, Female Life Expectancy	Number
le_raceadj_q4_F	Race-Adjusted, Q4, Female Life Expectancy	Number
le_agg_q4_F	Unadjusted, Q4, Female Life Expectancy	Number
le_raceadj_q1_M	Race-Adjusted, Q1, Male Life Expectancy	Number
le_agg_q1_M	Unadjusted, Q1, Male Life Expectancy	Number
le_raceadj_q2_M	Race-Adjusted, Q2, Male Life Expectancy	Number
le_agg_q2_M	Unadjusted, Q2, Male Life Expectancy	Number
le_raceadj_q3_M	Race-Adjusted, Q3, Male Life Expectancy	Number
le_agg_q3_M	Unadjusted, Q3, Male Life Expectancy	Number
le_raceadj_q4_M	Race-Adjusted, Q4, Male Life Expectancy	Number
le_agg_q4_M	Unadjusted, Q4, Male Life Expectancy	Number