CS498 Cloud Computing Applications

U.S. Health Data Analysis – Team 30

Nicolas Maire (nmaire2)

John Moran (jfmoran2)

Graham Chester (grahamc2)

MCS-DS

University of Illinois Urbana-Champaign

Champaign, USA

# Introduction

## Background and Motivation

The World Health Organization has declared reducing health inequities an important goal, and the United States has one of the least equitable health systems in the industrialized world [1]. There have been substantial investments in collecting data that can serve to inform and shape health policies that address existing inequities. While the primary reason for investing in such data collection was often not the study of health inequities, but rather of the rapid growth of health care spending [2], these datasets are nevertheless often disaggregated in a way that allows investigation of potential drivers of inequality.

In this project we are using the Hadoop ecosystem toolset to analyze United States health datasets as released by the Center for Disease Control and Prevention (CDC), and the Health Inequality Project. We 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 gender by major US City.

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 U.S., beyond the U.S., 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.

The intellectual merit of this study lies in developing a better understanding of health care inequalities and their impact on health in the United States today. Such understanding will be of broad interest as health and health care affects everyone equally.

## Related work

While the question of health care inequalities in industrialized countries has received considerable attention in recent years (see [3] for a review), studies published until recently were unable to conclusively identify the main societal drivers of health inequities (see for example [4]).

The 500 cities data set provides new opportunities for novel approaches, given the good-quality data at the appropriate spatial scale, and the data has very recently been used to study heterogeneity in the prevalence of some chronic conditions. An article linking obesity to places with low-income and minority populations was published while we were working on this project [5].

Given the considerable interest novel computational approaches and their potential for innovative data analytics have received recently due to success stories in many different application domains, one would expect a substantial body of literature on this topic in the field of evidence-based health research. However, as illustrated by a recent review article [6], there are to date few published studies or applications that employ cloud computing to address data analytical tasks in health research.

# Data

## Datasets

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 [6] 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 [7] provided by the Health Inequality Project (https://healthinequality.org), contains life expectancy by US city for men and women by income quartile.

## Data Wrangling

The CDC 500 cities health dataset is keyed by city, and the Health Inequality Project Life Expectancy by Income Dataset is keyed by commuting zone, so they cannot easily be joined, but being able to see both health and life expectancy by income is vital to our analysis.

A significant effort was required to research what data is available to allow these to be joined, and even more to source the data mapping files. The final solution required a file to map city to county code (FIPS), and another to map commuting zone to county code (FIPS), to enable the join.

At the completion of Data Wrangling we now have a cleansed dataset providing a comprehensive view of health, and life expectancy by income for 500 Cities across the United States.

## Initial Data Exploration

Tables of Top 10 most and least healthy places in to live were created using a (subjectively) weighted health score based on various Preventative Measures, Unhealthy Behaviors and Health Outcomes. We attempted to define these weights to align with public perception of the criticality of each of the measures.

The Spark MLlib statistics package was then used to generate a correlation matrix and heatmap as shown in the Preliminary Results section below.

# Cloud Infrastrucure and Toolset

We had proposed 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.

As described in the Progress to Date section, we have now refined this to be:

• Amazon EC2 Deep Learning AMU (Ubuntu) Instance

• Amazon S3 for data file storage

• Spark Dataframes, and Jupyter Notebooks using Python/Pyspark and the Seaborn library for data visualization.

• Spark MLlib for statistics and modelling

# Progress to date

## Challenges faced

### Infrastructure

Initially we installed Hortonworks HDP instance as a VirtualBox virtual machine (VM) on our individual PCs in order to test the required Hadoop components. However, we found that firstly there were components missing or not operational (e.g. unsupported old Python version, Zeppelin not operational, Jupyter not installed) that took significant time to resolve. Secondly the memory requirements were larger than our laptops could handle efficiently, and thirdly that it wasn’t a common instance for the team.

In order to resolve this our initial plan was to export our modified Hortonworks VirtualBox (VM), and create an Amazon Elastic Compute Cloud (EC2) instance from this. The AWS command line interface was used to upload the VM to an AWS Simple Storage Service (S3) bucket, and from there this was imported into EC2. Unfortunately, the AWS verification program reported that it was not compatible with EC2 due to an unsupported kernel version in the VM, and it was not clear if this could be resolved so an alternative plan was required.

### Toolset

The process of deciding on this toolset included investigating Hive as a query tool, by developing trial scripts (e.g. to find the best/worst cities for health score), we also developed a Java program to call Hive. However, it soon became clear that the interactive/programmatic statistical analysis we required would be infeasible using Hive.

We then investigated Pyspark with Zeppelin notebooks however Zeppelin on Hortonworks proved difficult to use and not properly configured.

Additionally, Spark shell and Scala proved to be time consuming to learn and had limited charting capabilities.

Further research into Spark revealed that for our structured tabular data that Spark RDD’s were not ideal and that Spark Dataframes would be a better option.

## Solutions

### Infrastructure

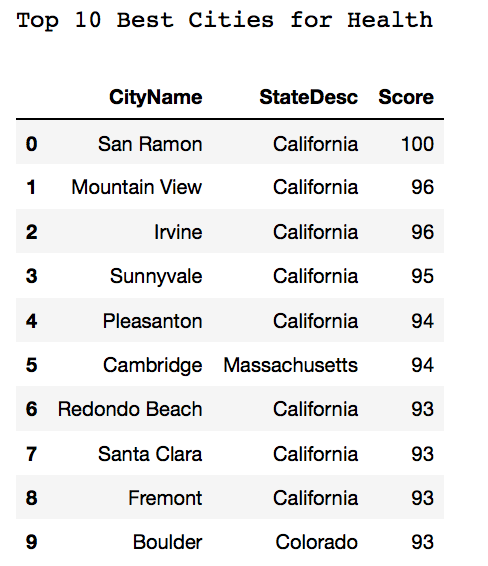
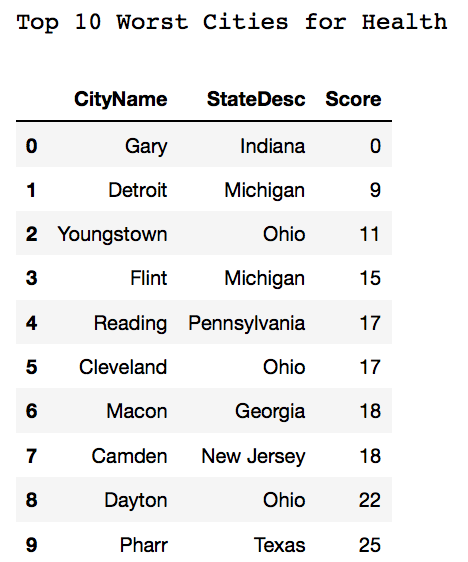
After trying unsuccessfully to resolve the unsupported kernel version error, we decided to look at pre-existing AMIs provided by AWS to see if there were ones that would support our PySpark/Jupyter notebook solution. We found an acceptable AMI, the Deep Learning AMI (Ubuntu) Version 5.0 AMI. With minimal configuration efforts we were able to upload our notebook and run it on this AMI. We uploaded all our data to an AWS S3 bucket and were able to successfully access the S3 bucket from the Jupyter notebooks running on our local browsers.

### Toolset

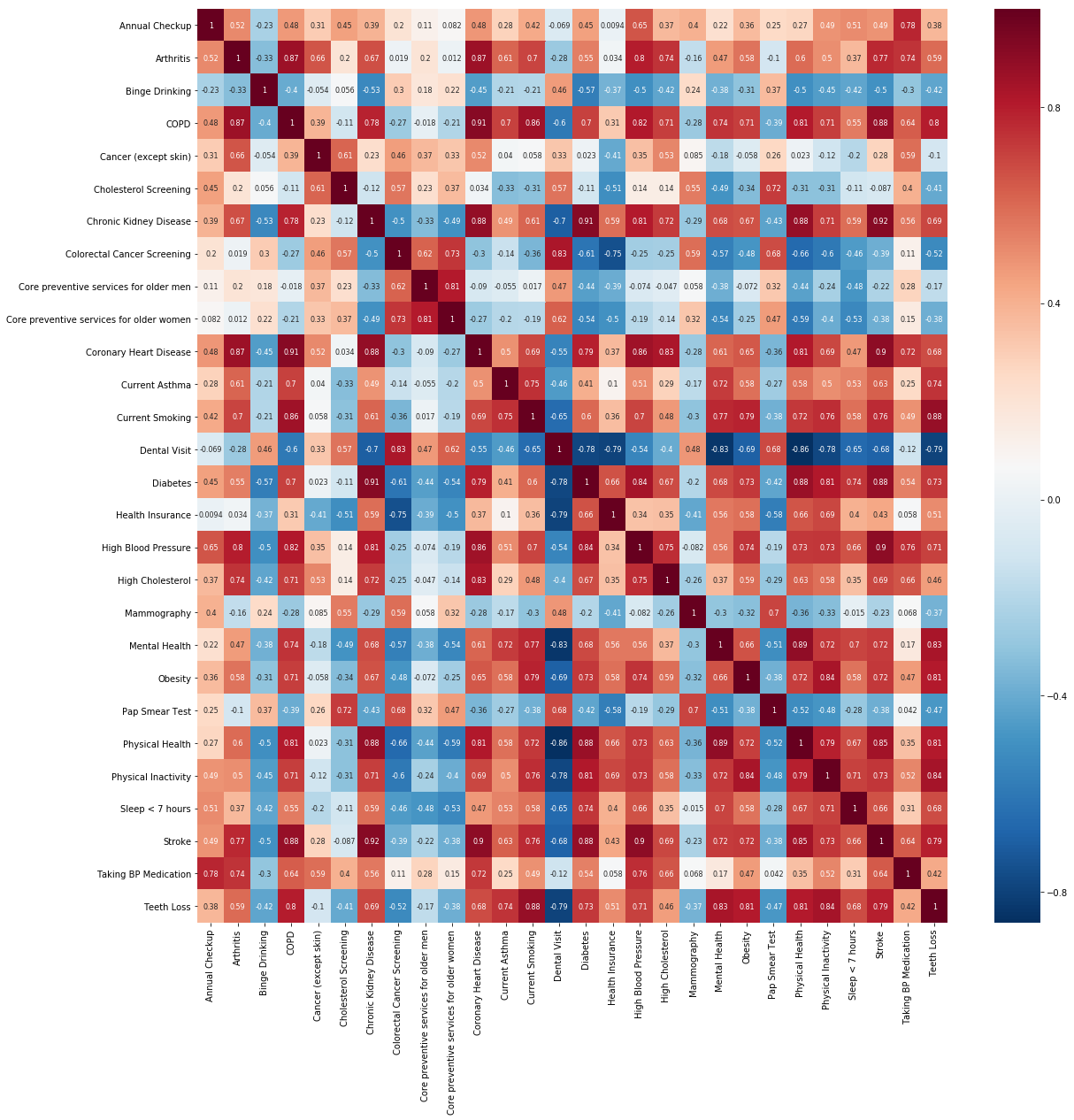
From the Challenges Faced section above we finally decided on Jupyter notebooks for the interactivity and flexibility, Spark Dataframes for high volume data capabilities and interactive performance, and PySpark to enable the use of Python and libraries such as Seaborn for graphing,

## Preliminary results

Table 0 shows the top 10 best and worst cities for health among the 500 cities analyzed.

The correlation heatmap below is an initial investigation of relationships between health measures. Some correlations indicate interesting and plausible interrelationships between health measures, while others are possible artefacts due to confounding factors, and will need further investigation.



# Discussion

## Prototype

A prototype version of the Jupyter notebook as run on the Amazon EC2 instance is provided as supplementary material (500citiesSpark.html).

## Future work – Cloud Infrastructure

We will investigate the level of effort required to run our current configuration on a cluster, as a proof of concept, but will first focus on improving and automating our current cloud services configuration.

## Future work – Spark Data Analyses

We have updated this section significantly from the project proposal as we developed a greater understanding of the potential of the available data.

We will also develop the Pyspark data analysis workbook further incorporating Spark MLlib and Statistics libraries as follows:

• model the impact of unhealthy behaviors on life expectancy for the wealth and the poor (top and bottom income quartiles).

• model the impact of serious disease on life expectancy for the wealth and the poor.

• model the benefit of preventative health measures on life expectancy for the wealth and the poor.

• model the impact of poverty on life expectancy

• examine the correlations between preventative health care and disease, and between unhealthy behaviors and disease.

## Timetable

Week 12-Investigate expanding AWS instance to a cluster

Week 13-Develop further analyses and models

Week 14-Continue to develop analyses and models

Week 15-Complete final write-up

Week 16-Final edit and submission

# Appendix

1. 500 Cities Dataset Definition

| 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 | City FIPS code | Plain Text |
| Measure | Measure full name | Plain Text |
| Data\_Value\_Unit | The unit,"%" for percent | Plain Text |
| DataValueTypeID | Id for the data value type | Plain Text |
| Data\_Value\_Type | 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 |

1. Health Risk Factors and Outcome as percent of population

|  |
| --- |
| Health Risk Factors |
| Binge drinking among adults aged >=18 Years |
| Doctor visit for routine checkup in the past year for adults aged >=18 Years |
| Cholesterol screening among adults aged >=18 Years |
| Fecal blood test, sigmoidoscopy, or colonoscopy, adults aged 50–75 Years |
| Older adult men aged >=65 Yrs up to date on a set of preventive services |
| Older adult women aged >=65 Yrs up to date on set of preventive services |
| 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 |
| High blood pressure control medication among adults aged >=18 Years |
| 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 in adults aged >=18 Years, 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 |

##### 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. Dieleman JL et al. US Spending on Personal Health Care and Public Health, 1996-2013. JAMA. 2016
3. A. Leigh, C. Jencks, T.M. Smeeding. Health and Economic Inequality. Oxford Handbook of Economic Inequality (2009), pp. 384-404
4. Wolfson M, Kaplan G, Lynch J, Ross N, Backlund E. Relation between income inequality and mortality: empirical demonstration. BMJ. 1999 Oct 9;319(7215):953-5.
5. Fitzpatrick KM, et al. Obesity and place: Chronic disease in the 500 largest U.S. cities. Obes Res Clin Pract (2018, in press)
6. Tawalbeh LA et al.: Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications (2016). IEE Access. Volume 4. 6171-6180.
7. 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>
8. Health Inequality (2017). Retrieved on 18-Feb-18 from <https://healthinequality.org/dl/health_ineq_all_online_tables.xlsx>