U.S. Health Data Analysis – Team 30

CS498 Cloud Computing Applications, 2018 Spring Semester

Nicolas Maire (nmaire2@illinois.edu)
John Moran (jfmoran2@illinois.edu)
Graham Chester (grahamc2@illinois.edu)
Department of Computer Science
University of Illinois Urbana-Champaign
Champaign, USA

***Abstract*—Reducing health inequality has been declared an important item on the public health agenda by the World Health Organization. We have studied health inequality across US cities and states by analyzing data on health services coverage and the prevalence of chronic diseases using a cloud architecture. We measured the performance of the cloud platform and investigated how it scales with the expected future growth of the data set using synthetic data sets.**

Our analysis revealed considerable differences between cities and states and offers insights into possible causal relationships. A ranking of cities based on a composite health score was created as a basis for guiding health system policy.

The cloud platform scaled with increasing data size and was shown to deliver satisfactory performance with a single-digit number of compute nodes for realistic real-world data size.

Keywords—Health, Life Expectancy, Inequality, Wealth, Poverty, CDC, Analysis, Spark, Cluster

I. INTRODUCTION

A. 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 care 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 the study 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 tools from the Hadoop ecosystem 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, published as "500 Cities: Local Data for Better Health" [3], to both identify the healthiest and unhealthiest places to live in the US, and to analyze key relationships between health risk factors and outcomes, and life expectancy and income by major US City.

The purpose of using a Hadoop cluster running in a cloud ecosystem 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. As additional datasets comparable to the 500 cities data are currently not publicly available, we have used a re-sampling approach to generate arbitrarily-sized synthetic datasets to investigate the scalability of our architecture.

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.

B. Related work

While the question of health care inequalities in industrialized countries has received considerable attention in recent years (see [4] for a review), studies published until recently were unable to conclusively identify the main societal drivers of health inequities (see for example [5]).

The CDC 500 cities data set provides an opportunity for novel approaches, given the good-quality data at the appropriate spatial scale, and the fact that this data has very recently been used to study heterogeneity in the prevalence of some chronic conditions. An article linking obesity to locales with low-income and minority populations was published while we were working on this project [6].

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. A workshop organized by the European Commission in 2014 on "High Performance Computing (HPC) in Health Research" concluded that there are considerable opportunities for the integration of HPC in health research [7]. However, as illustrated by a recent review article [8], there are, to date, few published studies or applications that employ cloud computing to address data analytical tasks in health research.

II. METHODS

A. DATA

1) Datasets

There are a broad variety of datasets from varying sources (and of varying quality) available regarding aspects of health in the United States. We chose these datasets firstly, because they are from renowned organizations, and secondly, because they 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), for 500 cities and census tracts within the United States.

The Health Inequality Dataset [9] provided by the Health Inequality Project (https://healthinequality.org), contains life expectancy by US city for men and women by income quartile.

2) 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. However, being able to see both health and life expectancy by income was vital to our analysis. It was a significant challenge to research what data is available to allow these to be joined, and then to source these 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 have a cleansed dataset providing a comprehensive view of health, and life expectancy by income for 500 Cities across the United States.

3) Endpoints of analysis

Tables of the Top 10 most and least healthy places in to live were created using a 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, in analogy to the disability weights used for the standard measurement of burden of disease [10]. We then used the Spark MLlib statistics package to generate a correlation matrix and heatmap as shown in the Results section below, and we used the Python Seaborn library for regression and charting.

B. Cloud Infrastrucure and Toolset

We used the following tools, mainly from the Hadoop ecosystem, to analyze the dataset:

- Amazon EC2 for Linux instances
- Amazon S3 for data and script storage
- Spark Datasets and Python/PySpark scripts for data wrangling
- Jupyter Notebooks with Spark Datasets, and Python/Pyspark for data analysis, with Seaborn for visualization.
- Spark MLlib for statistics and correlations.

C. Technologies Evaluated

1) 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.

After trying unsuccessfully to resolve an unsupported kernel version error, we decided to look at pre-existing AMIs provided by AWS to see if any would support our PySpark/Jupyter notebook solution. We found an acceptable AMI, the Deep Learning AMI (Ubuntu) Version 5.0 AMI. With minimal configuration effort we were able to upload our notebook and run it on this EC2 instance. We uploaded all our data to an AWS S3 bucket and, with the appropriate security in place, were able to successfully access the S3 bucket from the Jupyter notebooks running on our local browsers.

2) Toolset

The process of deciding on this toolset included investigating Hive as a query tool, developing trial scripts (e.g. to find the best/worst cities for health score), and developing 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 both 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 investigation 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.

We finally decided on Jupyter notebooks for the interactivity and flexibility, Spark Dataframes for its high-volume data capabilities and interactive performance, and PySpark to enable the use of Python and libraries such as Seaborn for charting,

D. Final Architecture

The cloud architecture for our project is shown in Figure 1 below:

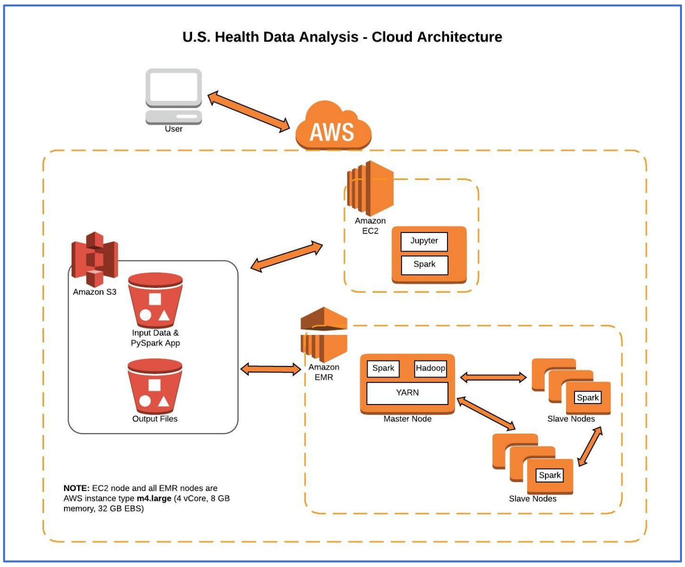

Figure 1

The system architecture consists of three main components:

1. AWS S3 Data Store
2. AWS EMR Cluster
3. AWS EC2 Node

The PySpark application was divided into two parts: 1) the data transformation that ran on an EMR cluster and 2) the data analysis and visualization that ran on an EC2 node. The input and output files for the application resided on an AWS S3 bucket.

1) AWS S3 Data Store

The input and output files shared a common Simple Storage Service (S3) bucket. The PySpark data transform application also resided on the same S3 bucket, so that no local EBS storage was utilized either on the EC2 node or EMR cluster.

2) AWS EMR Cluster

The PySpark data transformation ran as a standalone step on the EMR master node. The number of slave nodes was varied from 0 to 16 for our scalability testing. All EMR nodes were of AWS instance type m4.large with the following hardware characteristics: 4 Virtual Cores, 8 GB memory, 32 GB EBS storage. The default Ubuntu Deep Learning AMI was installed on each node, along with Hadoop 2.8.3, Spark 2.3.0, and Python 3.6.

The PySpark application ran as a YARN (Yet Another Resource Negotiator) client, and YARN and Spark handled the distribution of work over the EMR cluster. Spark, running on EMR, used the EMRFS (EMR File System), to directly access the input data stored on S3 and to write the output data to S3.

3) AWS EC2 Node

The Elastic Compute Cloud (EC2) node ran the Jupyter Notebook server using Python/PySpark. The EC2 node accessed the S3 data files on S3, that were previously generated by the EMR cluster, to perform data analysis using Spark MLib and data visualization using the Seaborn library.

The EC2 Node utilized the same node type as EMR, i.e. the AWS instance type m4.large. The Linux system configuration on the EC2 node was the Amazon Machine Instance (AMI) Deep Learning AMI (Ubuntu) Version 8.0.

The EC2 node could be accessed by any browser on the public internet that had access to the correct security key.

III. Results

A. Performance and Scalability

The Spark data transformation application was run on the master cluster node with Number of Cities = (500, 5000, 50000) and number of slave cluster nodes = (0, 1, 2, 4, 8, 16). The results are shown in Figure 2 below:

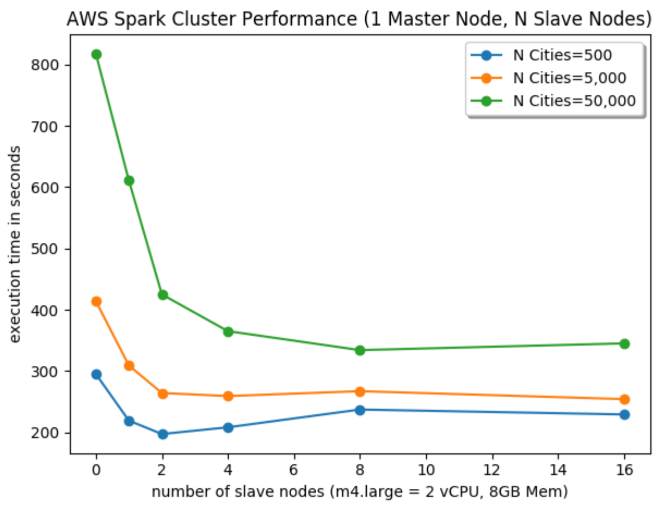

Figure 2

For each of our test sets of cities, we calculated the optimal number of slave nodes and the execution time improvement for that configuration. This data is shown in Table 1 below:

Number of Cities	Optimal Cluster Slaves Nodes	Execution time improvement
500	2	33%
5,000	4	37%
50,000	8	59%

Table 1

As an additional test, we ran a 500,000 cities instance. We did not include this in our results, as there are not 500,000 cities in the world, but we used it as a scaling check for optimal node calculation. It resulted in the same number of optimal slave nodes (8) as the 50,000 city test.

B. Data Analysis

1) Top 10 Cities

Table 2 shows the top 10 best and worst cities for overall health among the 500 cities analyzed.

Top 10 Best Cities for Health

	CityName	StateDesc	Score
0	San Ramon	California	100
1	Mountain View	California	96
2	Irvine	California	96
3	Sunnyvale	California	95
4	Pleasanton	California	94
5	Cambridge	Massachusetts	94
6	Redondo Beach	California	93
7	Santa Clara	California	93
8	Fremont	California	93
9	Boulder	Colorado	93

Top 10 Worst Cities for Health

	CityName	StateDesc	Score
0	Gary	Indiana	0
1	Detroit	Michigan	9
2	Youngstown	Ohio	11
3	Flint	Michigan	15
4	Reading	Pennsylvania	17
5	Cleveland	Ohio	17
6	Macon	Georgia	18
7	Camden	New Jersey	18
8	Dayton	Ohio	22
9	Pharr	Texas	25

Table 2

2) Best and Worst, States and Cities

In this and the following sections, we use the possibly emotive terms ‘rich’ to refer to the top 25% of income earners, and ‘poor’ to refer to the bottom 25%. The unhealthy behavior rate is an average of the rates for all unhealthy behaviors (e.g. smoking, obesity, etc.), and the same applies to the preventative measures rate (e.g. annual health check), and to the health outcomes rate (e.g. cancer, heart disease).

NB: Order is Best, 2nd Best, 3rd Best	**Best States**	**Worst States**
Life Expectancy	Hawaii, Idaho	Nevada, Oklahoma: ~2.8years worse
Rich/Poor Gap in Life Expectancy	California, New Mexico, Florida	Wyoming, Delaware, Maryland: ~5 years worse
Unhealthy Behavior Rate	Vermont, Colorado, Utah: ~21%	Mississippi, New Jersey, Delaware: ~30%
Serious Disease Rate	Vermont, North Dakota, Minnesota: ~10%	West Virginia, Ohio, Mississippi: ~16%
Heart Disease Rate	Vermont, DC, Alaska, Utah: ~4%	West Virginia, Ohio, Pennsylvania: ~8%
Smoking Rate	Utah, California, Hawaii: ~13%	Ohio, Maryland, Mississippi: ~25%
Obesity Rate	Vermont, Hawaii, Colorado: ~20%	Mississippi, Delaware, Louisiana: ~37%
Dental Health (>65yo, lost all teeth)	Hawaii, Minnesota, North Dakota: ~7%	Ohio, West Virginia, Mississippi: ~23%

NB: Order is Best, 2nd Best, 3rd Best	**Best Cities**	**Worst Cities**
Life Expectancy	Santa Fe NM, San Jose/Sunnyvale/Santa Clara CA	Columbus GA, Henderson NV, Las Vegas NV: ~4 years worse
Rich/Poor Gap in Life Expectancy	Laredo TX, Yuma AZ	Wichita Falls TX, Champaign IL, Decatour IL:~8 years worse
Unhealthy Behavior Rate	Newport Beach/Irvine CA, Boulder CO: ~17%	Detroit MI, Camden NJ, Youngtown OH, Flint MI, Gary IN: ~35%
Serious Disease Rate	College Station TX, Irvine/Mountain View CA: ~8%	Gary IN, Youngstown OH, Detroit/Flint MI: ~20%
Heart Disease Rate	College Station TX, Provo UT: ~2.6%	Youngstown OH, Gary IN: ~10%
Smoking Rate	Orem UT, Sunnyvale/Newport Beach CA: ~9%	Flint MI, Detroit MI, Youngstown OH: ~30%
Obesity Rate	Irvine/Fremont/Milpitas CA, Boulder CO: ~15%	Gary IN, Detroit MI, Reading PA, Birmingham AL: ~44%
Dental Health (>65yo, lost all teeth)	San Remon/Redondo Beach CA: ~5%	Gary IN, Trenton NJ, Youngstown OH: ~30%

Table 3

While, for brevity, Table 3 only shows the top two or three best/worst cities or states, it is notable that the same cities/states keep appearing in the best or worst of many measures. For example, Gary, Indiana is the worst city for dental health, obesity, smoking and is therefore among the worst cities for overall unhealthy behaviors. It is unlikely to be a coincidence that it also has the worst overall rate of Serious Disease, and the 5th lowest life expectancy. For states, Vermont has the lowest obesity rates (and lowest overall Unhealthy Behavior Rate), and again probably not coincidentally, the lowest serious disease rate.

Since states or cities that are the worst with one unhealthy behavior tend to be the worst with other unhealthy behaviors (and the same applies to serious diseases and preventative measures) there are clearly underlying probable socio-economic factor(s) driving this. Whilst there is insufficient data to establish the causal factors, we can hypothesize that income-related factors such as education and health-insurance may be significant contributing factors.

3) *Life Expectancy and Health*

We find a high level of stratification of life expectancy by income quartile, showing an approximate 8-year gap between the rich and the poor (Figure 3). We can also see the relationship between the general health of a city and life expectancy, though it appears to be a fairly weak relationship with only as little as a year for cities with the worst health compared with those with the best. Somewhat surprisingly, the impact of wealth on life expectancy is much more significant that the general health of the particular city.

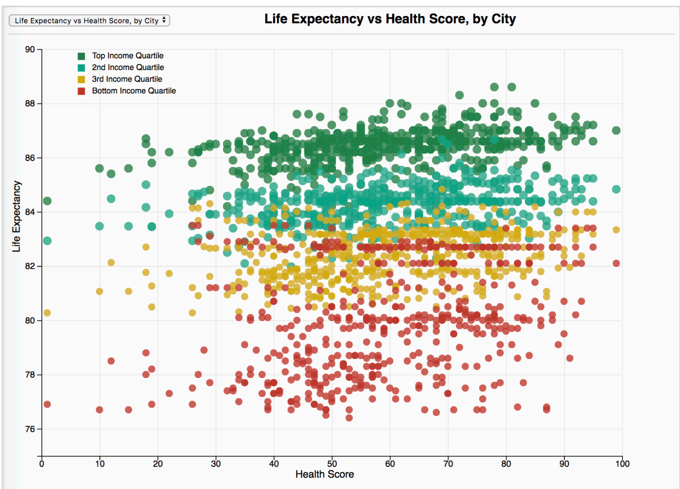

Figure 3

The following three charts' regression slopes show that life expectancy for the poor is much more adversely associated with Unhealthy Behaviours (Figure 4b), and Health Outcomes or Problems (Figure 4c), than for the rich. Figure 4a suggests that the poor gain much more benefit from preventative measures.

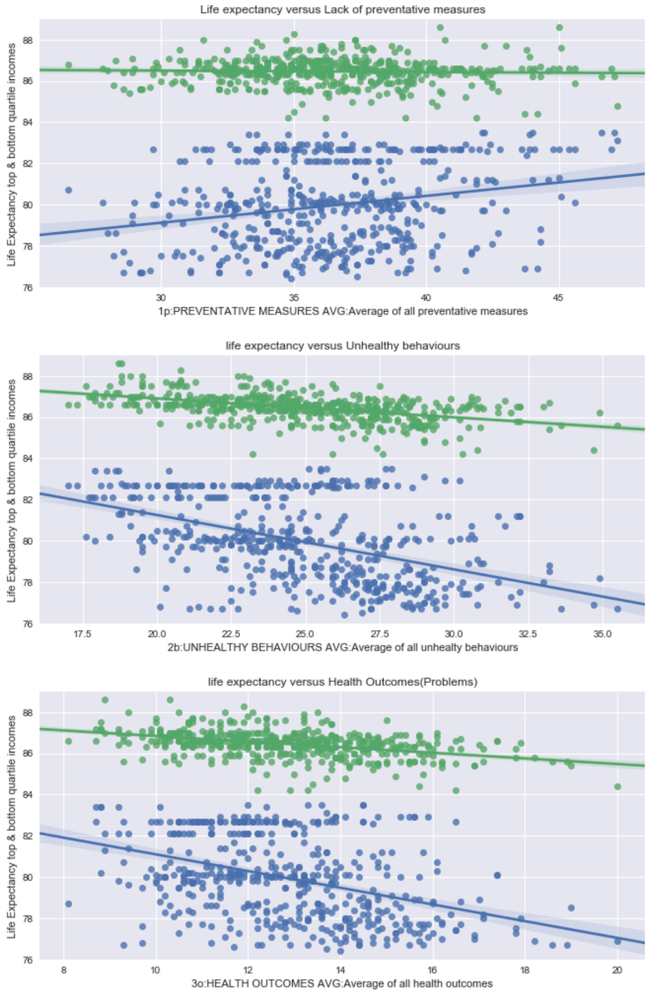

Figure 4

The following charts show that life expectancy for both the rich and the poor is affected by Smoking (Figure 5b), Obesity (Figure 5c) and Inactivity (Figure 5d), but again the regression slopes show that the poor are much more adversely affected than the rich. Somewhat surprisingly Binge Drinking seems to have almost no relationship to life expectancy, and for the rich there is actually a slight positive relationship (Figure 5a). We can only hypothesize that other factors such as level of education, being a college student where this is prevalent [11], or being wealthy enough to afford to binge drink is balancing the negative effects.

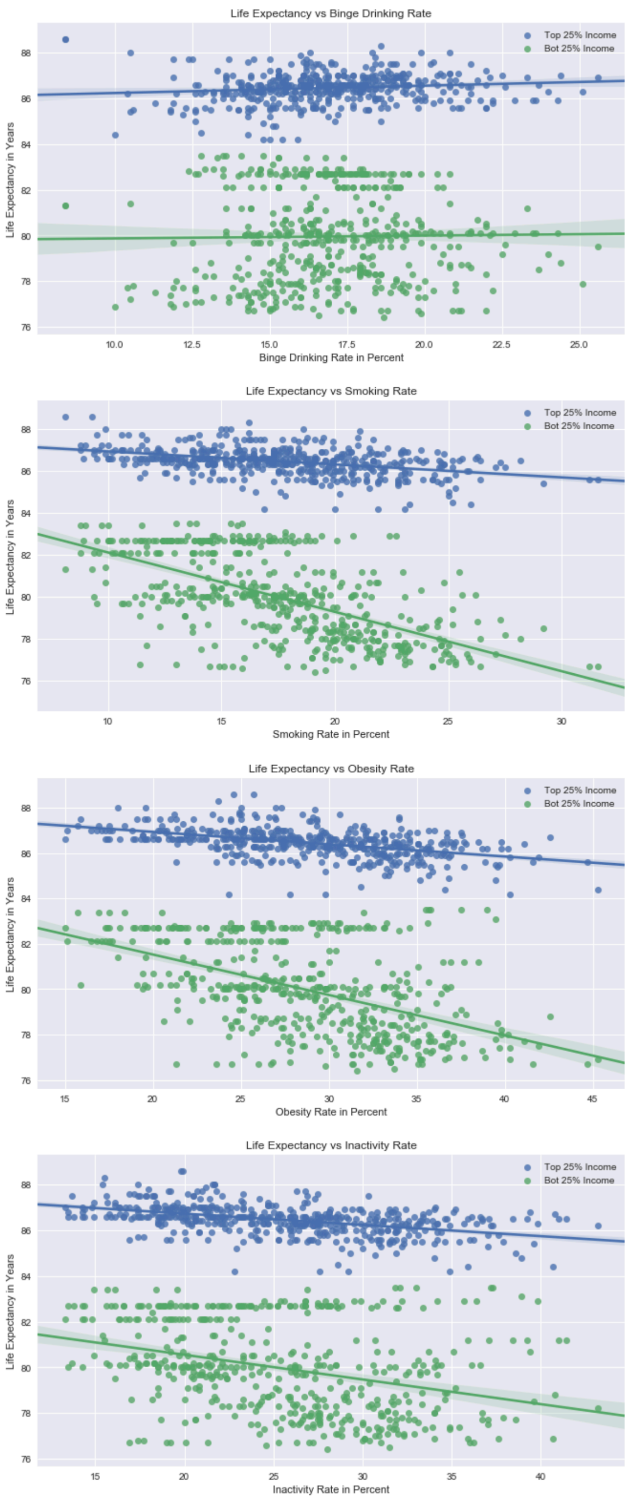

Figure 5

4) Other Notable Findings

Large cities (over 1 million) tend to have slightly worse preventative measures, and slightly worse unhealthy behaviors than smaller cities, However, the average life expectancy is essentially identical. Large cities also have slightly lower life expectancy gaps between rich and poor.

There is a region of central Illinois where the rich vs poor life expectancy gap is among the widest in the country (Champaign Illinois has the second-most unequal life expectancy between rich and poor). There is also region of far southern Texas where this gap is among the smallest.

There is a fairly high correlation (0.75) between average life expectancy and the gap in life expectancy between rich and poor. Unfortunately, it seems there is little trickle-down effect from good health in the rich to good health in the poor. This is confirmed by surprisingly low correlations of only 0.33 and 0.5, for women and men respectively, between the life expectancy for the rich and life expectancy for the poor.

In addition to the above, the correlation between Preventative Measures and Health Outcomes is only 0.27, so while there may be other hidden factors, this low rate does imply that Annual Checkups, Cholesterol screening etc. (full list in the Appendix, Table II), are having a limited impact on health outcomes (disease). Whereas the correlation coefficient of Unhealthy Behaviors and Health Outcomes is 0.84 indicating a very strong relationship between measures such as smoking obesity etc. and cancer, heart disease etc.

A reasonable (but not provable) hypothesis from this data is that this relationship is causative. So while Preventative Measures appear not to be particularly ‘preventative’, it appears that Unhealthy Behaviors may be strongly causative.

While the income vs life expectancy data showing a very significant impact of wealth (or lack thereof) on life expectancy, is bad news for those near the bottom of the socio-economic ladder. The good news is that this can to some extent be mitigated by an individual’s choice to reduce their Unhealthy Behaviors.

5) Correlations

The correlation heatmap below (Figure 6) 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.

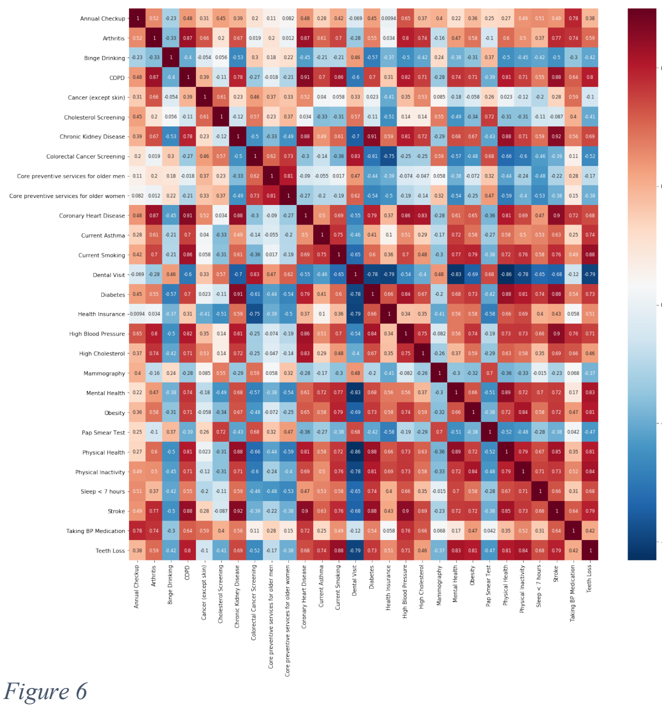

Figure 6

A closer look at the correlations between different measures reported in the data is largely consistent with expectation. For almost all pairs of unhealthy behaviors and preventive measures (combined as “risk factors” in the analysis), the sign of the correlation is negative for predictors generally associated with poor outcomes, and vice versa. A notable exception is binge drinking, which is negatively correlated with almost any other

Unhealthy Behavior measure as discussed in the Other Notable Findings section above.

Measures in the Health outcome category are mostly strongly positively correlated among themselves. Cancer is the only health problem which shows only small positive or even slightly negative correlations with other health outcomes.

An analysis of the relationship between risk factors (Unhealthy Behaviors) and health outcomes is significantly more problematic than the analysis within each of these two categories. The correlation structure is in many places hard to reconcile with well-established bio-medical facts: for example, there is no clear relationship between smoking and cancer apparent in the data. It is not surprising that the nature of the data doesn't capture the causal relationships in many cases: one obvious shortcoming is both risk factors and health outcomes are measured instantaneously, when in reality there is a substantial delay expected before health consequences manifest.

IV. CONCLUSIONS

Our analysis of a data set on health-relevant practices and health outcomes from 500 US cities using a cloud architecture confirmed that there are substantial health inequities at both the city and state levels. Our rankings of cities and states highlights the extremes as a first step to inform health policies towards a more equitable national health system.

Linking to additional data sets on wealth, and life expectancy, has corroborated the observations that poverty and poor health often go together, possibly because of the strong relationship of poverty and unhealthy behavior. In addition, the data set largely confirms that health inequities affect different areas of the health services and health outcomes in similar ways, but it also provides some interesting exceptions that would be worth exploring further.

Using a Spark AWS EMR cluster for this project proved to be an effective technology choice and scaled well as we increased the number of cities. Although the number of cities in the world is capped and finite, the dimensionality and complexity of our data analysis could increase in the future and we have shown that optimal performance can be achieved with a small number of cluster nodes.

V. APPENDIX

TABLE I. 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

Table 4

TABLE II. HEALTH RISK FACTORS AND OUTCOME AS PERCENT OF POPULATION

Unhealthy Behaviors
Binge drinking among adults aged >=18 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
Sleeping less than 7 hours among adults aged >=18 Years
Preventative Measures
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
Mammography use among women aged 50–74 Years
Papanicolaou smear use among adult women aged 21–65 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] 500 Cities: Local Data for Better Health 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

[4] A. Leigh, C. Jencks, T.M. Smeeding. Health and Economic Inequality. Oxford Handbook of Economic Inequality (2009), pp. 384-404

[5] 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.

[6] Fitzpatrick KM, et al. Obesity and place: Chronic disease in the 500 largest U.S. cities. Obes Res Clin Pract (2018, in press)

[7] European Commission: Workshop Report High Performance Computing in Health Research. Retrieved on 6 May from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=8642

[8] Tawalbeh LA et al.: Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications (2016). IEE Access. Volume 4. 6171-6180.

[9] Health Inequality (2017). Retrieved on 18-Feb-18 from https://healthinequality.org/dl/health_ineq_all_online_tables.xlsx

[10] Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. Lancet. 2012;380:2129–43.

[11] https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-facts-and-statistics