Semantic Web: Python Application to Interpret Data.gov Data

Thomas Sonderman, Lenore Montalbano

Florida Atlantic University

COP 5859

Contents

[Abstract 3](#_Toc425593412)

[Semantic Web: Python Application to Interpret Data.gov Data 4](#_Toc425593413)

[Data Sources 4](#_Toc425593414)

[Geographic Data 4](#_Toc425593415)

[Health Care Data 5](#_Toc425593416)

[Our Inquiry 6](#_Toc425593417)

[Methodology 6](#_Toc425593418)

[Analysis and Results 7](#_Toc425593419)

[References 8](#_Toc425593420)

[Tables 9](#_Toc425593421)

[Figures 12](#_Toc425593422)

[Python Code 13](#_Toc425593423)

Abstract

Do regions with suppressed economies have higher incidence of death? Does the death rate correspond with the medical services being offered in these economically suppressed areas? These are the questions we have attempted to answer in this paper through the use of the Semantic Web. We obtained our raw data via Geonames.org and Data.gov, through the use of DCAT we were able to assemble a data store, by utilizing languages such as Python, pandas and numpy we were able to query the data store and derive a result set. Based upon the result set we could then run queries against them and derive an answer to our questions. In essence we discovered that the poverty level within a state does not dictate the death rate for that state, however the states with a higher number of hospitals have a lower death rate.

Semantic Web: Python Application to Interpret Data.gov Data

The objective of the Semantic Web is to establish well-defined meaning or semantics for information on the web, enabling better cooperation between computers and people. Over the years the road traveled to define and shape Semantics has evolved and many tools have been created that allows for the establishment of a defined semantic; plus the ability to combine and filter data has grown.

The objective for this project is to develop an application that utilizes GeoNames.org as a primary source of data in conjunction with additional sources of data to perform statistical analysis on this data.

This paper will begin with an overview of the resources used to obtain data, the questions being posed, the tools used to answer these questions, code samples, result sets and finally an analysis of the results.

# Data Sources

Our primary concern was securing data that was fair and accurate; by accessing public information stores we felt that the data available would not be subject to corporate influence or favoritism.

## Geographic Data

We obtained our geographic information from GeoNames.org, this is a geographic database which is available for download under a creative commons attribution license, meaning it is free of charge, although donations are welcome. The data accessible through the various web services and data exports contain over 10 million geographical names and consists of over 9 million unique features with 2.8 million named place and 5.5 million alternate names. It is through GeoNames.org that we obtained information regarding regions throughout the United States.

## Health Care Data

Our interest covered several aspects of health related information, our main source of data was pulled through Data.gov and related websites. Data.gov is a United States government web site with a stated purpose of increasing public access to high value, machine readable datasets that have been generated by the Executive Branch of the Federal Government. The data available on this site is not private nor restricted for national security reasons.

**Hospital Data**

The data downloaded from data.medicare.gov 3 provides a look into all the hospital data that is available on data.gov. The data provides location, name, and affiliation information. This data will give us a map of the healthcare availability per region/location.

**Death Data**

The data downloaded from data.cdc.gov 2 provides access to a wealth of information on deaths with location and death type data, this will map the same geographic locations of deaths with the hospital availability.

**CHSI Indicators**

The indicators downloaded from [ftp.cdc.gov](ftp://ftp.cdc.gov) 1 are provided by location to look at health risks. The CHSI indicators can provide an insight into the risk levels that are associated with the geographic regions we have previously mapped.

## Our Inquiry

The output of this data will answer some questions that we laid out. Are we able to look at a specific region and see if they offer decent hospital access but have a high number of deaths? If the deaths are high, is this consistent with the regions CHSI indicators or poverty rate? These questions if answered correctly could provide a useful correlation between the health data on data.gov and human behavior. If there are risks, are they confined to specific regions of the United States?

One thing we will be looking at is whether there are any specific trends in the data. Do large metropolitan areas have higher health risks/deaths? With this data we can explore many different hypotheses about health and geographic locations.

## Methodology

To analyze the data gleaned from the sources identified above an application with the ability to access those sources, combine the data and query against that data providing a resulting data set has been developed using the Python programming language.

The use of Python, pandas and numpy enable us to create datasets from the csv files we created via our download. DCAT is a standard that allowed us to combine portions of the datasets and generate tabular output for further analysis.

The enclosed code demonstrates the strategy employed to access the data sources, combine the data sources and generate output based upon specific requests. We grouped our inquiry based upon states and generated a table that provided the total number of hospitals, poverty level, population total and death totals per state, see Table 1.

## Analysis and Results

Our first analysis was to determine the an overall median for death rate, poverty level and health care providers via hospitals, having benchmarked those three indicators we were now free to query the data. We asked the same question three different ways:

1. Do the states with the highest death rates have the highest poverty levels or the fewest hospitals?
2. Do the states with the highest poverty levels have the fewest hospitals and the highest death rates?
3. Do the states with the most hospitals have the lowest death rates?

We sorted the data based upon the first criteria in each question and determined the median of each of our data points, Table 2 shows the accumulation of this data and Figure 1 displays a graphical analysis of the results.

What we discovered is that although a state has the highest death rate for their population they do not have the highest poverty level, not do they have the fewest hospitals, and this high number of deaths may be attributed to another factor which we didn’t contemplate in our analysis. Surprisingly the states with the highest poverty level neither had the highest death rates nor the fewest hospitals, poverty in not indicator of high death rates in this analysis. However it should be noted that the states with the most hospitals did have the fewest deaths but poverty level played no real part in the death rate nor the availability of health care.

In conclusion, by utilizing the tools made available due to the implementation of the Semantic Web our project has demonstrated the ability to ask areal world questions. Through the various public websites we were able to access standardized data and through the use of the Python programming language and additional libraries we were able to combine data in such a way that allowed us to create datasets that we could then query against.

Our inquiry resulted in an analysis that demonstrated the minimal impact of a high poverty level in overall death rates with in the United States, but we also discovered that those states with the most hospitals did possess the lowest median death rates overall.

References

1. CHSI Indicators. (n.d.). In ftp.cdc.gov online website. Retrieved from ftp://[ftp.cdc.gov/pub/Health\_Statistics/NCHS/Datasets/CHDI/chsi\_dataset.zip](http://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/CHDI/chsi_dataset.zip)
2. Death Data. (n.d.). In data.cdc.gov online website. Retrieved from <https://data.cdc.gov/api/views/qpap-3u8w/rows.csv?accessType=DOWNLOAD>
3. Geographic Data (n.d.) In geonames.org online website. Retrieved from http://www.geonames.org
4. Hospital Data. (n.d.). In data.medicare.gov online website. Retrieved from <https://data.medicare.gov/api/views/xubh-q36u/rows.csv?accessType=DOWNLOAD>

Tables

Table 1

Table 1: Overall State Totals

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| State | Death Ttls | Hospital Ttls | State Pop | Death % | Hospital % | Poverty % |
| Alabama | 17,733 | 91 | 4,849,377 | 3.66 | 1.88 | 18.9 |
| Alaska | - | 21 | 736,732 | 0.00 | 0.00 | 10.1 |
| American Samoa | - | 1 | - | 0.00 | 0.00 | 0 |
| Arizona | 16,532 | 79 | 6,731,484 | 2.46 | 1.17 | 18.6 |
| Arkansas | 3,795 | 78 | 2,966,369 | 1.28 | 2.63 | 19.4 |
| California | 66,126 | 339 | 38,802,500 | 1.70 | 0.87 | 16.8 |
| Colorado | 10,864 | 79 | 5,355,866 | 2.03 | 1.48 | 12.9 |
| Connecticut | 7,776 | 31 | 3,596,677 | 2.16 | 0.86 | 10.7 |
| Delaware | 607 | 7 | 935,614 | 0.65 | 0.75 | 12.9 |
| District of Columbia | - | 9 | 658,893 | 0.00 | 0.00 | 18.8 |
| Florida | 20,760 | 188 | 19,893,297 | 1.04 | 0.95 | 17.1 |
| Georgia | 9,881 | 134 | 10,097,343 | 0.98 | 1.33 | 19 |
| Guam | - | 1 | - | 0.00 | 0.00 | 0 |
| Hawaii | 4,178 | 18 | - | 0.00 | 0.00 | 11.2 |
| Idaho | 3,283 | 42 | 1,634,464 | 2.01 | 2.57 | 15.6 |
| Illinois | 18,851 | 181 | 12,880,580 | 1.46 | 1.41 | 14.6 |
| Indiana | 20,543 | 127 | 6,596,855 | 3.11 | 1.93 | 15.8 |
| Iowa | 1,544 | 118 | 3,107,126 | 0.50 | 3.80 | 12.6 |
| Kansas | 6,362 | 133 | - | 0.00 | 0.00 | 13.9 |
| Kentucky | 4,107 | 95 | 4,413,457 | 0.93 | 2.15 | 18.8 |
| Louisiana | 11,971 | 120 | 4,649,676 | 2.57 | 2.58 | 20 |
| Maine | - | 35 | 1,330,089 | 0.00 | 0.00 | 14 |
| Maryland | 7,172 | 47 | 5,976,407 | 1.20 | 0.79 | 10.2 |
| Massachusetts | 17,423 | 67 | 6,745,408 | 2.58 | 0.99 | 11.9 |
| Michigan | 11,095 | 136 | 9,909,877 | 1.12 | 1.37 | 17 |
| Minnesota | 8,257 | 132 | 5,457,173 | 1.51 | 2.42 | 11.2 |
| Mississippi | - | 97 | - | 0.00 | 0.00 | 23.9 |
| Missouri | 10,727 | 108 | 6,063,589 | 1.77 | 1.78 | 15.8 |
| Montana | - | 58 | 1,023,579 | 0.00 | 0.00 | 16.1 |
| National | - | - | - | 0.00 | 0.00 | 0 |
| Nebraska | 4,951 | 90 | 1,881,503 | 2.63 | 4.78 | 12.9 |
| Nevada | 15,879 | 34 | 2,839,099 | 5.59 | 1.20 | 15.8 |
| New Hampshire | - | 26 | 1,326,813 | 0.00 | 0.00 | 9 |
| New Jersey | 6,656 | 64 | 8,938,175 | 0.74 | 0.72 | 11.4 |
| New Mexico | 6,794 | 42 | 2,085,572 | 3.26 | 2.01 | 21.4 |
| New York | 76,127 | 180 | 19,746,227 | 3.86 | 0.91 | 16 |
| North Carolina | 9,886 | 111 | 9,943,964 | 0.99 | 1.12 | 17.8 |
| North Dakota | - | 45 | 739,482 | 0.00 | 0.00 | 11.6 |
| Northern Mariana Islands | - | 1 | - | 0.00 | 0.00 | 0 |
| Ohio | 49,668 | 170 | 11,594,163 | 4.28 | 1.47 | 15.9 |
| Oklahoma | 4,747 | 127 | 3,878,051 | 1.22 | 3.27 | 16.7 |
| Oregon | 6,491 | 61 | 3,970,239 | 1.63 | 1.54 | 16.5 |
| Pennsylvania | 7,574 | 174 | 12,787,209 | 0.59 | 1.36 | 13.7 |
| Puerto Rico | - | 52 | - | 0.00 | 0.00 | 0 |
| Rhode Island | 3,260 | 12 | 1,055,173 | 3.09 | 1.14 | 14.7 |
| South Carolina | - | 63 | 4,832,482 | 0.00 | 0.00 | 18.5 |
| South Dakota | - | 55 | 853,175 | 0.00 | 0.00 | 14 |
| Tennessee | 29,316 | 111 | 6,549,352 | 4.48 | 1.69 | 17.8 |
| Texas | 60,209 | 374 | 26,956,958 | 2.23 | 1.39 | 17.5 |
| Utah | 9,206 | 46 | 2,942,902 | 3.13 | 1.56 | 12.6 |
| Vermont | - | 15 | 626,562 | 0.00 | 0.00 | 12.3 |
| Virgin Islands | - | 2 | - | 0.00 | 0.00 | 0 |
| Virginia | 5,693 | 86 | 8,326,289 | 0.68 | 1.03 | 11.7 |
| Washington | 15,919 | 92 | 7,061,530 | 2.25 | 1.30 | 14.1 |
| West Virginia | - | 52 | 1,850,326 | 0.00 | 0.00 | 18.4 |
| Wisconsin | 3,976 | 127 | 5,757,564 | 0.69 | 2.21 | 13.5 |
| Wyoming | - | 29 | 584,153 | 0.00 | 0.00 | 10.9 |

Table 2

Table 2: Top 10 Highest

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Poverty % | Death % | Hospital % | Population | Hospitals | Deaths |
| Overall Median: | 14.1 | 0.99 | 1.03 | 3,596,677 | 55 | 5,786 |
| Deaths: | 1.75 | 2.46 | 0.60 | 299,463 | 36 | 11,021 |
| Poverty: | 4.75 | 1.35 | 0.75 | 2,102,688 | 49 | 5,143 |
| Hospitals: | 2.05 | 0.40 | 1.54 | -104,089 | 52 | (1,359) |

Note: Numbers below the overall median depict the difference between the established benchmark and the top 10 highest states median values

Figures

Figure 1

Figure 1 Compare Median to Top 10 Highest Results

Python Code

matplotlib inline

import numpy as np

import math

import pandas as pd

import matplotlib.pyplot as plt

states = {

'AK': 'Alaska',

'AL': 'Alabama',

'AR': 'Arkansas',

'AS': 'American Samoa',

'AZ': 'Arizona',

'CA': 'California',

'CO': 'Colorado',

'CT': 'Connecticut',

'DC': 'District of Columbia',

'DE': 'Delaware',

'FL': 'Florida',

'GA': 'Georgia',

'GU': 'Guam',

'HI': 'Hawaii',

'IA': 'Iowa',

'ID': 'Idaho',

'IL': 'Illinois',

'IN': 'Indiana',

'KS': 'Kansas',

'KY': 'Kentucky',

'LA': 'Louisiana',

'MA': 'Massachusetts',

'MD': 'Maryland',

'ME': 'Maine',

'MI': 'Michigan',

'MN': 'Minnesota',

'MO': 'Missouri',

'MP': 'Northern Mariana Islands',

'MS': 'Mississippi',

'MT': 'Montana',

'NA': 'National',

'NC': 'North Carolina',

'ND': 'North Dakota',

'NE': 'Nebraska',

'NH': 'New Hampshire',

'NJ': 'New Jersey',

'NM': 'New Mexico',

'NV': 'Nevada',

'NY': 'New York',

'OH': 'Ohio',

'OK': 'Oklahoma',

'OR': 'Oregon',

'PA': 'Pennsylvania',

'PR': 'Puerto Rico',

'RI': 'Rhode Island',

'SC': 'South Carolina',

'SD': 'South Dakota',

'TN': 'Tennessee',

'TX': 'Texas',

'UT': 'Utah',

'VA': 'Virginia',

'VI': 'Virgin Islands',

'VT': 'Vermont',

'WA': 'Washington',

'WI': 'Wisconsin',

'WV': 'West Virginia',

'WY': 'Wyoming'

}

columns = ['state','deaths']

index = np.arange(57) # array of numbers for the number of samples

df = pd.DataFrame(columns=columns, index = index)

headers = ['Area', 'Year','Deaths']

df = pd.read\_csv('deaths.csv',header=False,

names=headers)

df['Area'] = df['Area'].astype(str)

df['Area'] = df['Area'].str[-2:]

byState\_df = pd.DataFrame(states.items())

byState\_df['Deaths'] = 0

byState\_df['Hospitals'] = 0

byState\_df['StatePop'] = 0

byState\_df['DeathPercentage'] = 0.0

byState\_df['HospitalPercentage'] = 0.0

byState\_df['PovertyPercentage'] = 0.0

df['Deaths'] = df['Deaths'].convert\_objects(convert\_numeric=True)

#print byState\_df

for index, row in df.iterrows():

if row['Area'] in states:

if not math.isnan(row['Deaths']):

byState\_df.loc[byState\_df[0] == row['Area'], 'Deaths'] = byState\_df.loc[byState\_df[0] == row['Area'], 'Deaths'] + row['Deaths']

#print byState\_df

headers = ['Provider ID','Hospital Name','Address','City','State','ZIP Code','County Name','Phone Number','Hospital Type','Hospital Ownership','Emergency Services','Location']

hospitals\_df = pd.read\_csv('hospitals.csv',header=False,

names=headers)

#print hospitals\_df['State']

for index, row in hospitals\_df.iterrows():

if row['State'] in states:

byState\_df.loc[byState\_df[0] == row['State'], 'Hospitals'] = byState\_df.loc[byState\_df[0] == row['State'], 'Hospitals'] + 1

#print byState\_df

headers = ['SUMLEV','REGION','DIVISION','STATE','NAME','POPESTIMATE2014','POPEST18PLUS2014','PCNT\_POPEST18PLUS']

statepop\_df = pd.read\_csv('state\_populations.csv',header=False,

names=headers)

#print hospitals\_df['State']

for index, row in statepop\_df.iterrows():

if row['NAME'] in states:

byState\_df.loc[byState\_df[0] == row['NAME'], 'StatePop'] = row['POPESTIMATE2014']

#print byState\_df

for index, row in byState\_df.iterrows():

if row['Deaths'] > 0 and row['StatePop'] > 0 and row['Hospitals'] > 0:

byState\_df.loc[byState\_df[0] == row[0], 'DeathPercentage'] = 1000 \* float(row['Deaths']) / float(row['StatePop'])

byState\_df.loc[byState\_df[0] == row[0], 'HospitalPercentage'] = 100000 \* float(row['Hospitals']) / float(row['StatePop'])

headers = ['StateCode','State','Poverty']

statepop\_df = pd.read\_csv('poverty.csv',header=False,

names=headers)

for index, row in statepop\_df.iterrows():

if row['State'] in states:

byState\_df.loc[byState\_df[0] == row['State'], 'PovertyPercentage'] = row['Poverty']

print byState\_df

byState\_df.to\_csv('output.csv')