Effects of Education, Home Health care, Poverty and Unemployment on people's Life Expectancy

Team Name : *Semantico*
Member 1 : *Abhishek Jajal (APJ180001)*
Member 2 : *Praveen Ramani (PXR170005)*

Project Type: *Custom*

1. INTRODUCTION

Human health and well being is a topic of great interest to a variety of organizations. In this project we focus on one aspect of this, which is life expectancy, which can be an indicator of the health of a society and level of prosperity. We plot life expectancy against several factors and derive the correlation along with some other statistics, for each comparison.

We first preprocess the data using R studio, to remove empty and irrelevant fields and then the data is converted into rdf using our custom software. We then run our own fuseki server to create a sparql endpoint locally and then the processed data is loaded into the sparql endpoint. We then fetch data from this, using ajax calls with sparql queries, from our website and then present the processed data using google api.

2. TARGET AUDIENCE

This project has a wide variety of interested parties like:

- The government, which can use this to identify issues in society and make informed policy decision.
  - e.g. Funding hospitals and health care professionals.

- Private organizations like health insurance companies, pharmaceutical companies etc., which can use this to identify business opportunities and create business plans.
  - e.g. Insurance companies can adjust their premiums and coverages.

- Individuals like investors, social workers, citizens etc., who can decide on the actions to be taken and amount of resources to allocate.
  - e.g. Social workers can volunteer assistance and raise funding.

3. DESCRIPTION OF DATA SOURCES

We have used 5 data sources for this project. All the datasets are in csv format and so downloaded as such. The data is converted into rdf after the preprocessing step. They are as follows:

3.1 DATASETS:

3.1.1 Poverty Dataset:

URL: https://catalog.data.gov/dataset/county-level-data-sets/resource/280dff75-cace-458a-bc4d-ff7c67a8366c

Description:

This dataset contains the poverty rates, in percentage, of all the states. It includes the state-wise poverty rates of the total population and that of the children population, along with the lower and upper bounds of the 90% confidence interval.

Pre-Processing:

We consider only the states and their corresponding poverty rate of the total population and the rest of the columns are dropped.

3.1.2 Unemployment Dataset:

URL: https://catalog.data.gov/dataset/county-level-data-sets/resource/6117e794-f5f6-47b0-90d1-ab32272595b1

Description:

This dataset contains state-wise unemployment rate, in percentage, from 2010 to 2018 and also the median household income.

Pre-Processing:

We consider only the states and their corresponding unemployment for the year 2015, and the rest of the columns are dropped.

3.1.3 Education Dataset:

URL: https://catalog.data.gov/dataset/county-level-data-sets/resource/c098e7fa-c9bd-40c1-a1f4-d13486cbd9eb

Description:

This dataset contains state-wise college completion rates, in percentage, including that of total, rural and urban areas, for the years 1970, 1980, 1990, 2000 and 2013-2017.

Pre-Processing:

We consider only the states and their corresponding total college completion rate for the years 2013-2017, and the rest of the columns are dropped.

3.1.4 Life Expectancy Dataset:

URL: https://catalog.data.gov/dataset/u-s-life-expectancy-at-birth-by-state-and-census-tract-2017

Description:

This dataset contains county-wise data of the life expectancy at birth along with it’s range and error, for the duration of 2010-2015.

Pre-Processing:

The county-wise data is converted into state-wise data in order to make it compatible with the other datasets. This is done by finding the average of the count-wise values of each state. The other columns are dropped.

3.1.5 Home Health Care Dataset:

URL: https://catalog.data.gov/dataset/home-health-care-state-by-state-data-5b494/resource/68058f90-ca0e-42fc-86c4-ddd6065c8cd9

Description:

This dataset contains state-wise data of various home health care statistics such as star rating, how often they got flu vaccines etc.,

Pre-Processing:

We consider only the state-wise star rating of patient care and the rest of the columns are dropped. The states were depicted using state abbreviations, which we manually changed to their respective state names in order to make it compatible with the other datasets.

3.2 DATA PREPROCESSING:

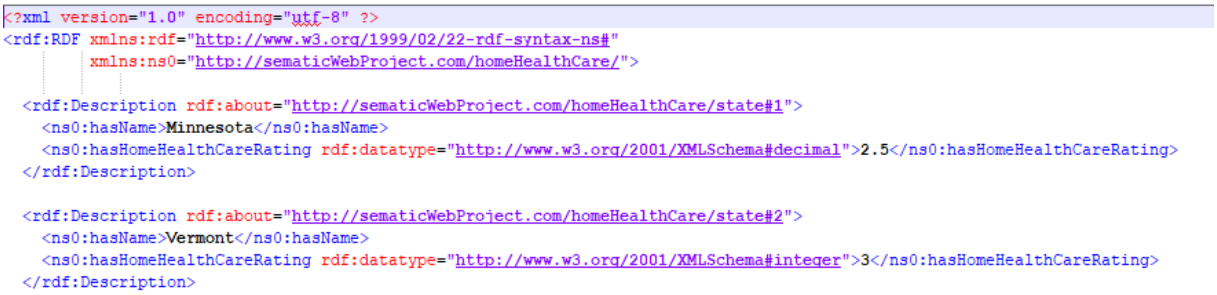
We have used R studio to do all the pre processing. After the initial preprocessing, which removes all the irrelevant columns and empty values in the dataset, we check all the pair wise correlations of all the data.

	Avg_Life_Expectancy	**Education**	**Rating**	**Poverty**	**unemployment**
Avg_Life_Expectancy	1.0000000	0.48411619	-0.3895727	-0.7949739	-0.49147639
Education	0.48411619	1.0000000	-0.2270897	-0.4852021	-0.08091264
Rating	-0.3895727	-0.22708973	1.0000000	0.3161289	0.12814482
Poverty	-0.7949739	-0.4852021	0.3161289	1.00000000	0.61792907
unemployment	-0.49147639	-0.08091264	0.12814482	0.6179291	1.00000000

Table 1. The pair-wise correlations of all the attributes

3.3 DATA CONVERSION:

All the preprocessed datasets are in csv format and is converted into turtle format using a custom parser. The turtle files are then converted into rdf files using this online tool: http://www.easyrdf.org/converter



```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:ns0="http://sematicWebProject.com/homeHealthCare/">

  <rdf:Description rdf:about="http://sematicWebProject.com/homeHealthCare/state#1">
    <ns0:hasName>Minnesota</ns0:hasName>
    <ns0:hasHomeHealthCareRating rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">2.5</ns0:hasHomeHealthCareRating>
  </rdf:Description>

  <rdf:Description rdf:about="http://sematicWebProject.com/homeHealthCare/state#2">
    <ns0:hasName>Vermont</ns0:hasName>
    <ns0:hasHomeHealthCareRating rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">3</ns0:hasHomeHealthCareRating>
  </rdf:Description>
```

Fig 1. Illustration of the semantic structure of the homehealthcare.rdf dataset

Here every row in the table has been converted into a resource with column name as predicate and the cell value as the object.

4. DATA INTEGRATION

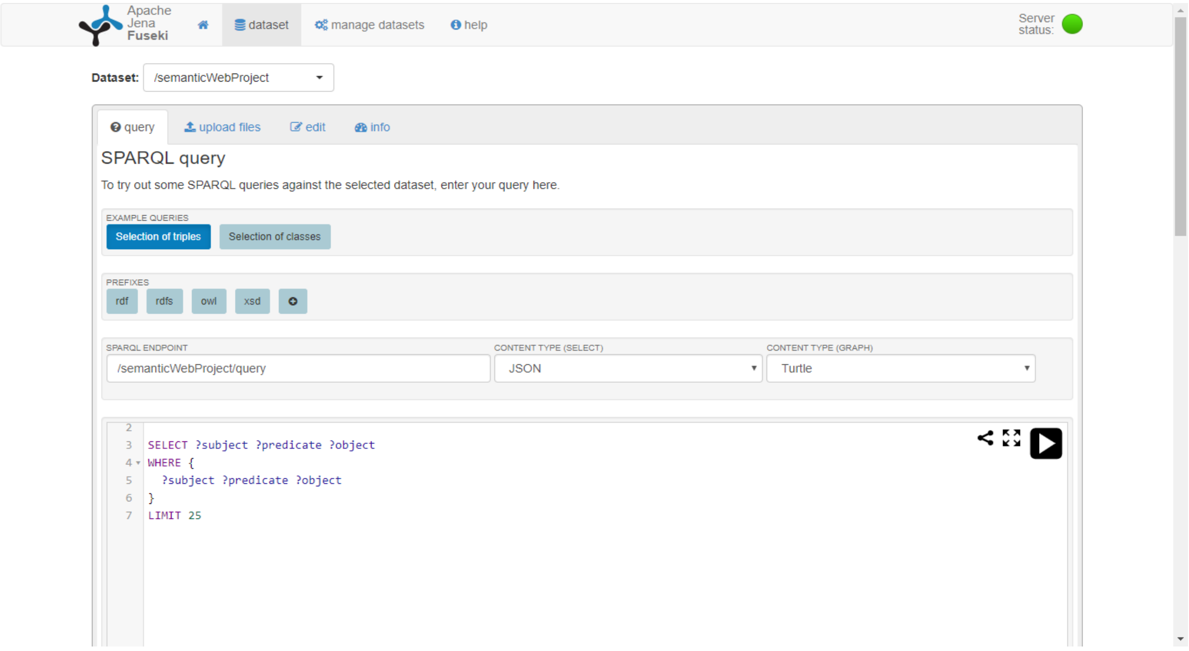
We first create a custom sparql endpoint using Fuseki server and load in the preprocessed rdf data. We then create a website which integrates the data and displays the results in a presentable manner using google api.

4.1 CUSTOM SPARQL ENDPOINT:

The fuseki server is run on our local machine and a dataset by the name of 'semanticWebProject' is created. The 5 rdf data files are then uploaded into this custom sparql endpoint and is ready to be queried as illustrated by Fig 2.

4.2 WEBSITE:

A website is created, which queries the data from this sparql endpoint and displays the results in a graphical manner using google api. An ajax call, containing a sparql query, is made to the custom sparql endpoint. The results of the query is fetched in json format and is presented in a graphical manner using google api libraries.



Dataset: /semanticWebProject

SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

SPARQL ENDPOINT: /semanticWebProject/query

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```

SELECT ?subject ?predicate ?object
WHERE {
  ?subject ?predicate ?object
}
LIMIT 25
```

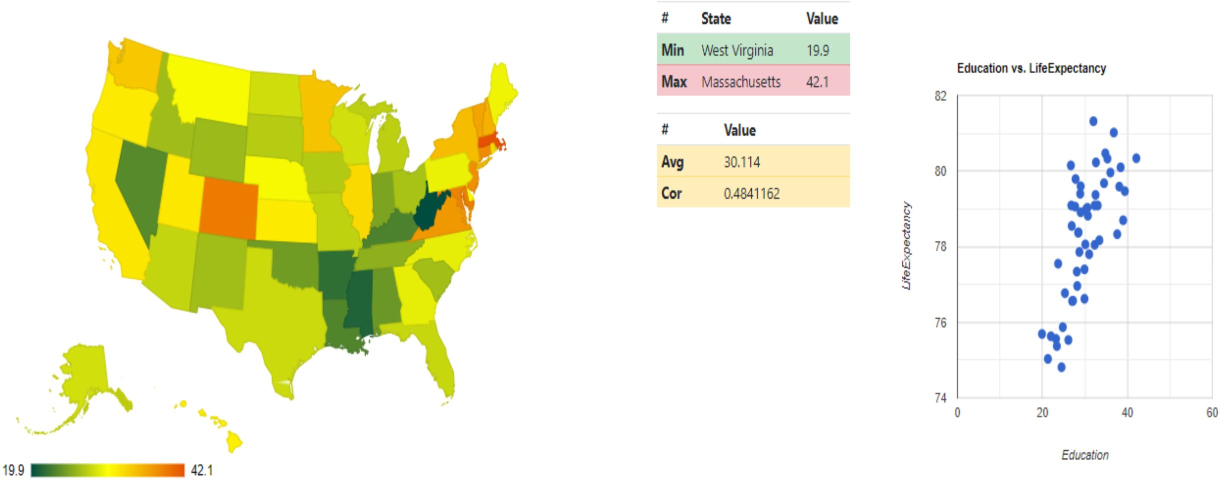
Fig 2. Custom SPARQL endpoint 'semanticWebProject'

5. DATA PRODUCT RESULTS

We have found several interesting results which show good correlation with life expectancy and are as follows:

5.1 Education Vs Life Expectancy:

We have found a good positive correlation value of 0.48 between these two datasets, showing that the more educated a state is, the higher is their life expectancy. We also found that the state with the least education is West Virginia with 19.9% and the one with the most is Massachusetts with 42.1%. The national average is 30.11%.



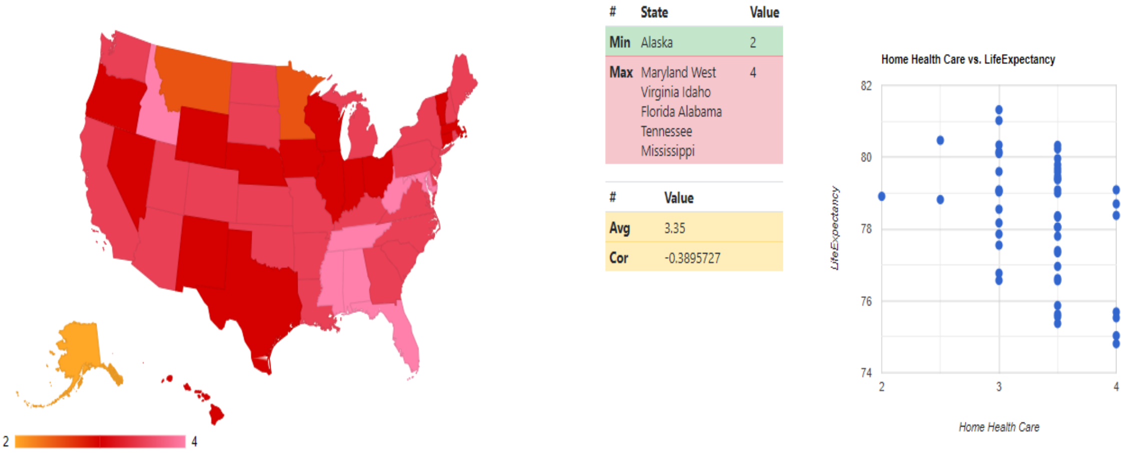
#	State	Value
Min	West Virginia	19.9
Max	Massachusetts	42.1

#	Value
Avg	30.114
Cor	0.4841162

Fig 3. Education Vs Life Expectancy

5.2 Home health care Vs Life Expectancy:

We have found a negative correlation of -0.389 between home health care rating and life expectancy, which is very unexpected and counter intuitive but the reason for this is that the home health care rating pertains only to the elderly while the life expectancy dataset pertains to the state wide population. Therefore, this result cannot be considered.



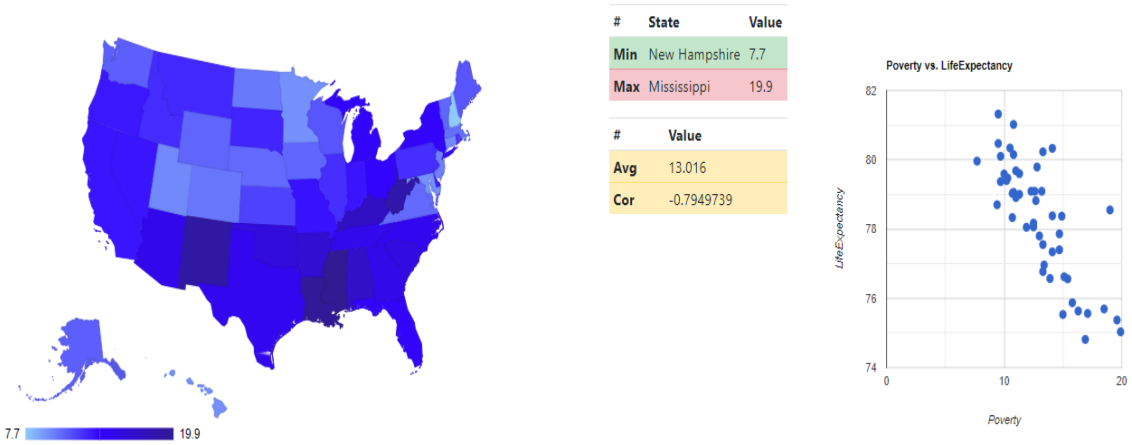
#	State	Value
Min	Alaska	2
Max	Maryland West Virginia Idaho Florida Alabama Tennessee Mississippi	4

#	Value
Avg	3.35
Cor	-0.3895727

Fig 4. Home health care Vs Life Expectancy

5.3 Poverty Vs Life Expectancy:

We have found a very good negative correlation of -0.79 between poverty and life expectancy, showing that the states with higher poverty rates have a lower life expectancy. We also found that New Hampshire has the lowest poverty rate at 7.7% while Mississippi has the highest at 19.9%. The national average is 13.01%.



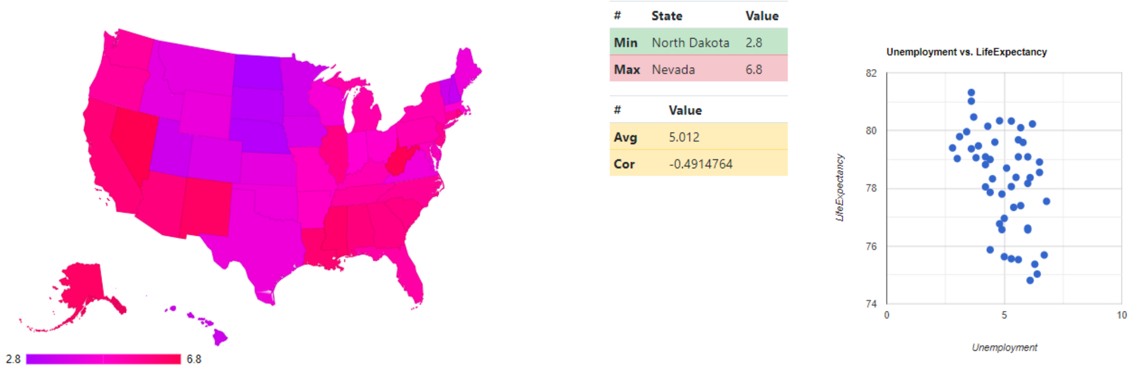
#	State	Value
Min	New Hampshire	7.7
Max	Mississippi	19.9

#	Value
Avg	13.016
Cor	-0.7949739

Fig 5. Poverty Vs Life Expectancy

5.4 Unemployment Vs Life Expectancy:

We have found a good negative correlation of -0.49 between unemployment and life expectancy, showing that the states with higher unemployment rates have lower life expectancy. We also found that North Dakota has the minimum unemployment rate of 2.8% while Nevada has the highest at 6.8%. The national average is 5.012%.



#	State	Value
Min	North Dakota	2.8
Max	Nevada	6.8

#	Value
Avg	5.012
Cor	-0.4914764

Fig 6. Unemployment Vs Life Expectancy

6. CUSTOM PROJECT JUSTIFICATION AND SUMMARY

We decided to change our project to a custom project due to various challenges that we encountered and to get some substantial results.

6.1 Justification:

Our project is a custom project for the following reasons:

6.1.1 Datasets:

Our project uses 5 different datasets while a simple project requires about 2 or 3 datasets. We do this to get a broader picture of the factors that effect life expectancy and to build a stronger case for our project.

The data was available only in csv formats and so we built a custom program to convert it into turtle and this was then converted into rdf using an online tool. In contrast, a simple project makes use of readily available rdf data.

6.1.2 SPARQL Endpoint:

Our project uses a custom sparql endpoint while a simple project makes use of a publicly available sparql endpoint. We do this by running the Fuseki server on our local machine, creating a custom endpoint and loading the rdf data into it.

6.1.3 Data processing:

Due to the large number of datasets and data attributes, we used R studio to preprocess the data in order to remove unnecessary data attributes, data holes and to find the correlation between various attributes in order to find some interesting results. A simple project does not involve preprocessing of data.

6.2 Summary:

This project processes unstructured data using semantic web technologies in order to extract some useful information from the datasets. The results are fetched from the custom sparql endpoint using an ajax call and the results are displayed on the website in a graphical manner using google api.

We have also derived some statistics like correlation between datasets and have also discussed the relevance and the implications of the correlation, presented in the website, to the target audience who would find it useful.