PROJECT REPORT

CS 6315 SEMANTIC WEB (F18)

Exercise for a healthy heart (Custom)

Team Extreme: Hanvitha Gavini, Prathyusha Kanmanth Reddy

Table of Contents
1. Introduction
2. Target Audience
3. Description of Data Sources
4. Data Integration
5. Results
6. Custom Project Justification
7. Summary

1. INTRODUCTION

The main aim of the project is to demonstrate the relationship between cardiovascular health and maintaining a healthy life. We analyzed both the Heart Disease and the Physical Activity data for each state and demonstrated the relationship between the two, based on ethnicity. We have used two datasets in this project, one gives the information about the Heart Disease and other contains information about the Physical Activity rate.

1.1. TOOLS USED

- Jena Fuseki Server
- Google Visualization API
- HTML/Bootstrap/ Java Script
- Sgvizler library

Jena Fuseki server is used as a SPARQL endpoint since there was no endpoint for the datasets used on the data.gov website. Using the Google Visualization API and Sgvizler library, the queried results are rendered to the webpage using JavaScript, HTML and Bootstrap.

Our project demonstrates results with two visualizations, Geo Map and Line Chart. The results on both the visualizations are based on ethnicity.

The Geo Map demonstrates the Heart Disease and Physical Activity distribution for each state on the United States Map. We can get the statistics of a state by hovering the mouse over the respective state region.

The Line Chart demonstrates the pattern or correlation between the Heart Disease and Physical Activity which helps us realize how they are closely connected.

2. TARGET AUDIENCE

- This project can be helpful for the Healthcare Communities and agencies to observe the relationship between cardiovascular health and regular physical activity.
- The results achieved can be used to demonstrate to the people to help them understand the benefits of exercising on their overall health.
- Also, the individuals can use the results as a proof to motivate them to hit the gym regularly.

3. DESCRIPTION OF DATA SOURCES

3.1 Behavioral Risk Factor Data: Heart Disease & Stroke Prevention (2011 to 2015)

- Source:https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Behavioral-Risk-Factor-Data-Heart-Disease-Stroke-P/4ny5-qn3w
- Format: RDF file
- Number of Triples: 2,440,350
- Number of Entities in the Data Set: 27

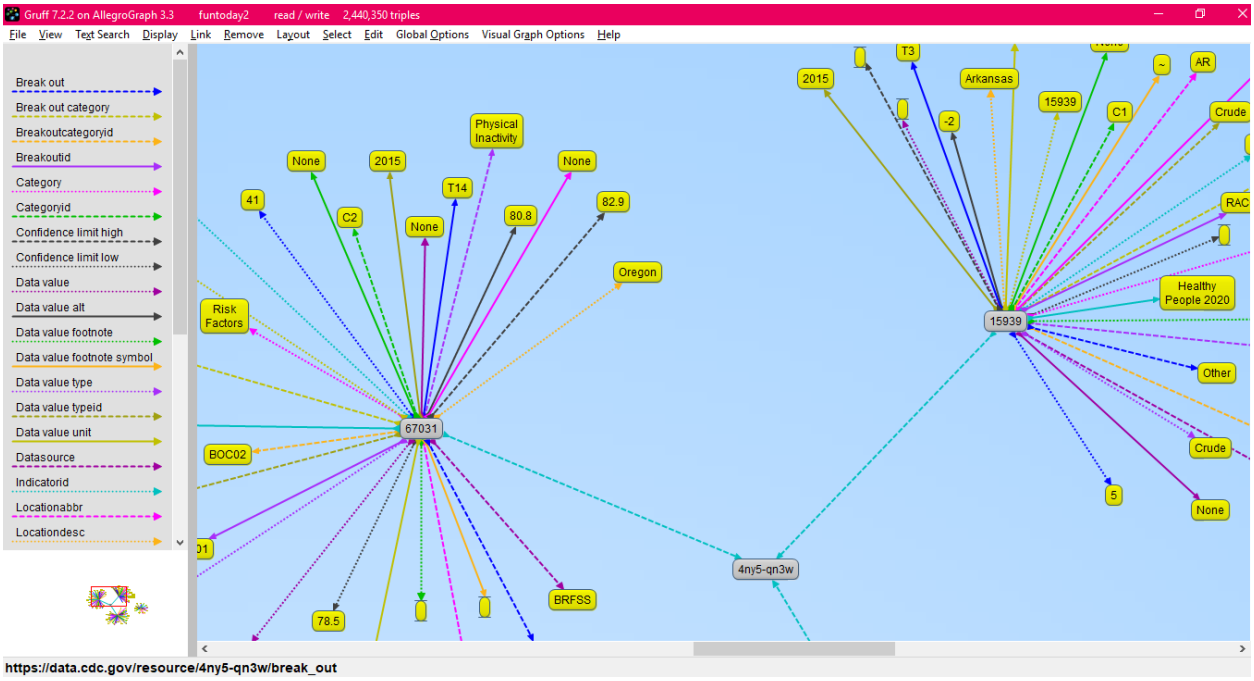

Figure a: Heart Disease & Stroke Prevention dataset ontology represented using GRUFF

3.2 Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System

- Source:https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7
- Format: RDF file
- Number of Triples: 1,655,152
- Number of Entities in the Data Set: 29

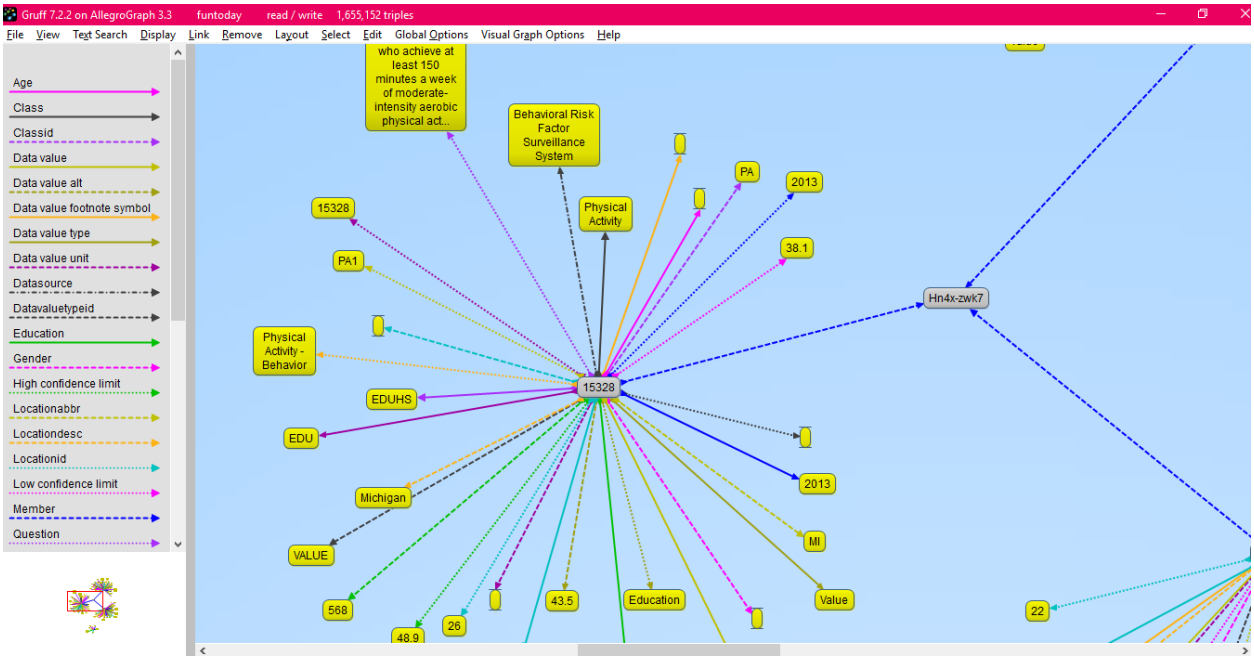

Figure b: Nutrition, Physical Activity, and Obesity dataset ontology represented using GRUFF

4. DATA INTEGRATION

- The two datasets are integrated using the common attribute "locationabbr", which is the abbreviation for representing various states in the United States of America.

- We wrote different SPARQL Queries, to achieve both Heart Disease and Physical Activity values for each state based on the ethnicity. For heart disease dataset, we set

    - LocationAbbr property as "State"
    - Break_Out_Category property is used for race
    - IndicatorId, BreakOutId to filter on race
    - And got back State's name and its corresponding Data Value

  For Physical Activity dataset, we set

    - YearStart property less than 2016 to match Dataset 1
    - LocationAbbr property as "State"
    - StratificationCategoryId1 for filtering on race
    - StratificationId1 is used for particular race code
    - Class to filter out 'Physical Activity' data
    - And got back State's name and its corresponding Data Value

- We used Jena Fuseki's Custom SPARQL endpoint to upload, integrate and query over the two datasets.

- The results from the SPARQL query are then passed onto the Google Visualization API.

5. RESULTS

The final results that are obtained after mashing up the two datasets based on ethnicities are demonstrated on a webpage using Sgvizler library and Google's Visualization API. It can be seen from the results that there is a strong correlation between heart disease diseases and physical activity especially for ethnic groups of White and Hispanic. On the other hand, for ethnic groups like Black and Asians, there is a weak correlation.

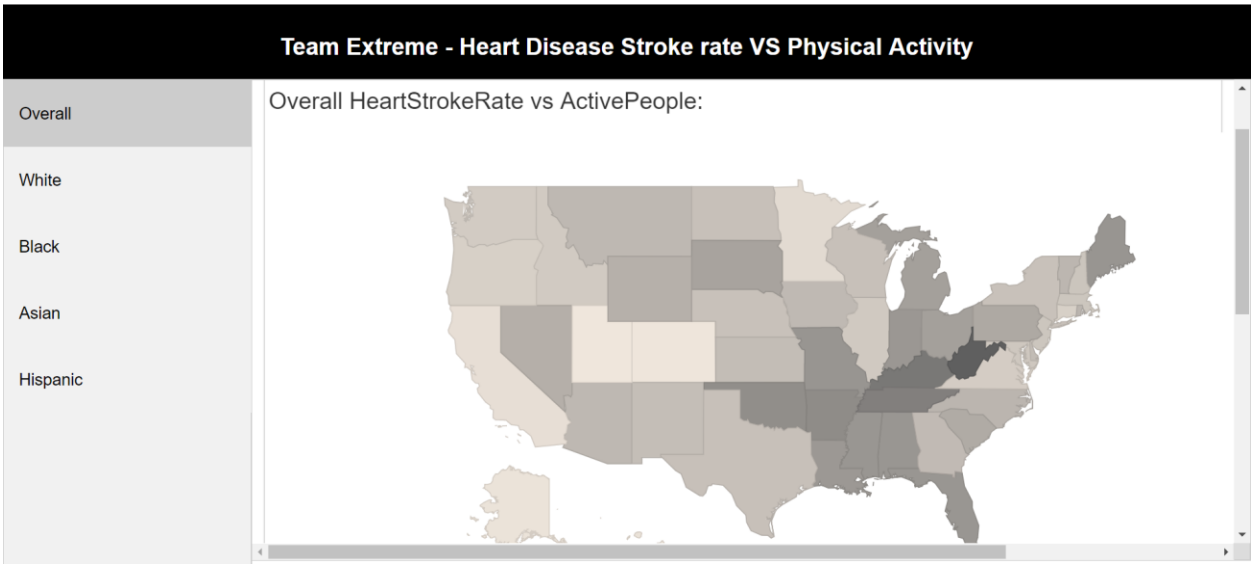

Figure c: Website's Homepage with project information and option tabs for user to make a selection

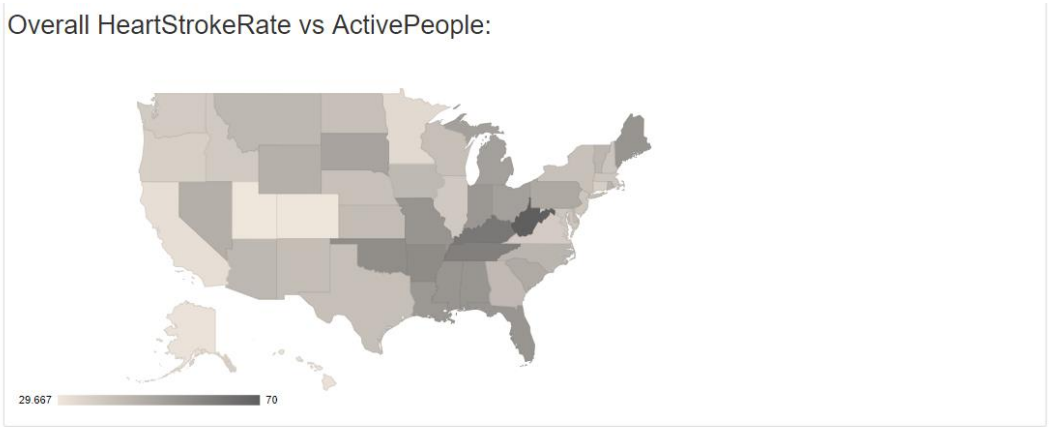

Figure d: GeoMap representation of Heart Disease and Physical Activity. User can see these statistics by simply hovering over the respective state.

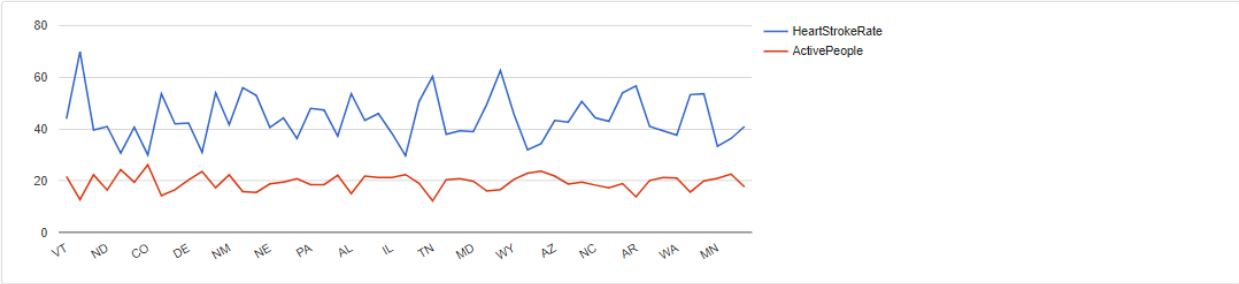

Figure e: Line Chart representation of Heart Disease and Physical Activity. User can see these statistics by simply hovering over the respective state.

6. CUSTOM PROJECT JUSTIFICATION

- SPARQL endpoint was setup on our local machines using Jena's FUSEKI server since there were no available endpoints on the web.
- For conversion of CSV files into RDF files, we tried almost all the available RDF converters to accomplish this task and by writing our own java API, we were able to achieve our goal.

7. SUMMARY

The goal of this project is to use Semantic Web Technologies to visualize the benefits of physical activity on a person's cardiovascular health. The resulting visualizations clearly shows that regular physical activity has direct impact on the health of the heart and states having the highest heart disease rate can also be seen.