PROJECT REPORT

SEMANTIC WEB (S17)
CS -6315

# Heart or Tobacco – It's your choice!
# (Custom Project)

By

Team Mavericks

Muhammad A Malik (mxm162431)
Teja Kiran Chunduri (txc163430)

# 1. INTRODUCTION

The main aim of the project is to demonstrate the adverse effect of the usage of tobacco on cardiovascular health. We analyzed both the Heart Mortality rate and the Average Tobacco use 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 Mortality rate and other contains information about the Average Tobacco use.

## 1.1. TOOLS USED

- Jena Fuseki Server
- RDF123
- Google Visualization API
- HTML/Bootstrap/ Java Script
- Sgvizler 0.6 library

Jena Fuseki server is used as a SAPRQL endpoint since there was no endpoint for the datasets used on the data.gov website. The CSV files were converted into RDF files using RDF123. 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 Mortality rate and Average Tobacco Use distribution for each state on the United States Map. We can get Heart Mortality rate and Average Tobacco use statistics of a state by hovering the mouse over the respective state region.

The Line Chart demonstrates the pattern between the Heart Mortality rate and Average Tobacco use, which helps us realize how tobacco usage is effecting the heart mortality rate.

# 2. TARGET AUDIENCE

- This project can be helpful for the Healthcare Communities and agencies to observe the effect of tobacco use over the cardiovascular health.
- The results achieved can be used to demonstrate to the people to help them understand the adverse effects of this terrible habit on their overall health.
- Also, the individuals can use the results as a proof to help someone close to them understand the harmful effects of tobacco.

# 3. DESCRIPTION OF DATA SOURCES

## **3.1** Behavioral Risk Factor Data: Tobacco Use (2011 to present)

- Source: https://catalog.data.gov/dataset/behavioral-risk-factor-data-tobacco-use-2011-to-present
- Format: RDF file
- Number of Triples: 177,575
- Number of Entities in the Data Set: 15

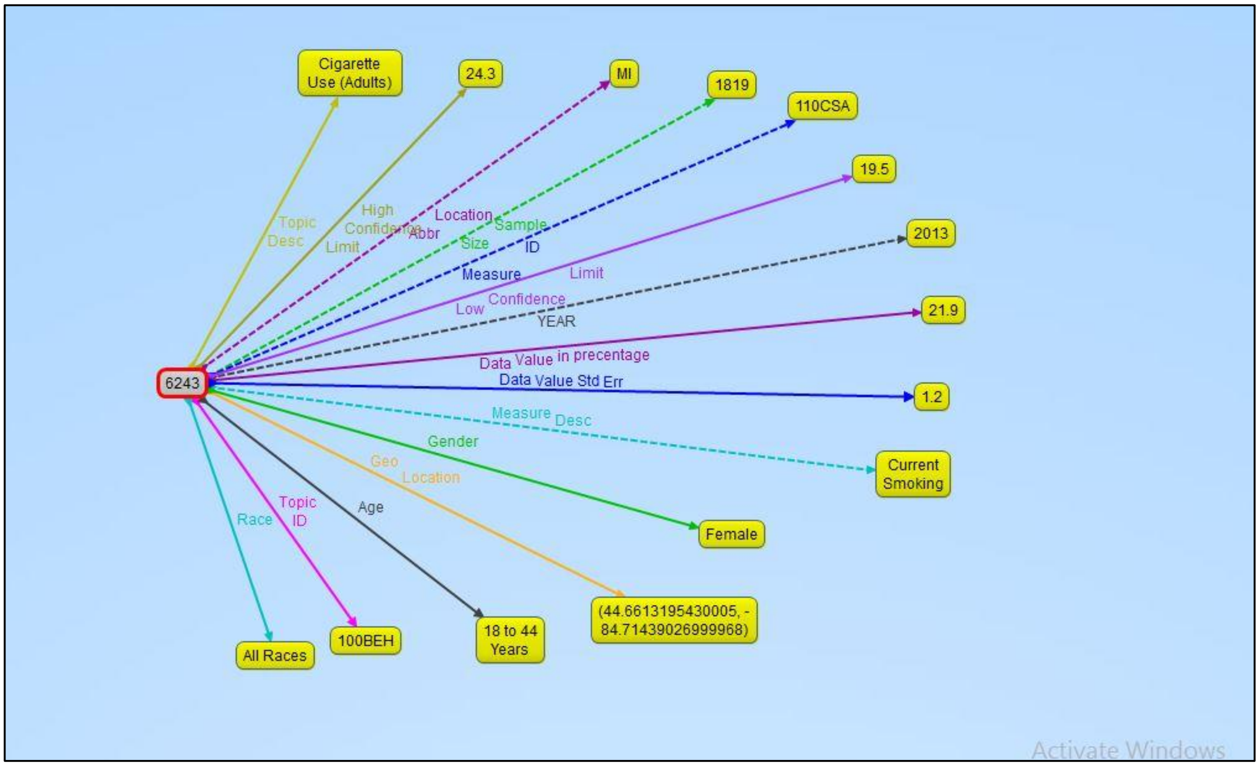


*Figure a: Tobacco Use dataset ontology represented using GRUFF*

## **3.2** Heart Disease Mortality Data among US Adults (35+) by State/Territory and County

- Source: https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county
- Format: RDF file
- Number of Triples: 1,013,938
- Number of Entities in the Data Set: 9

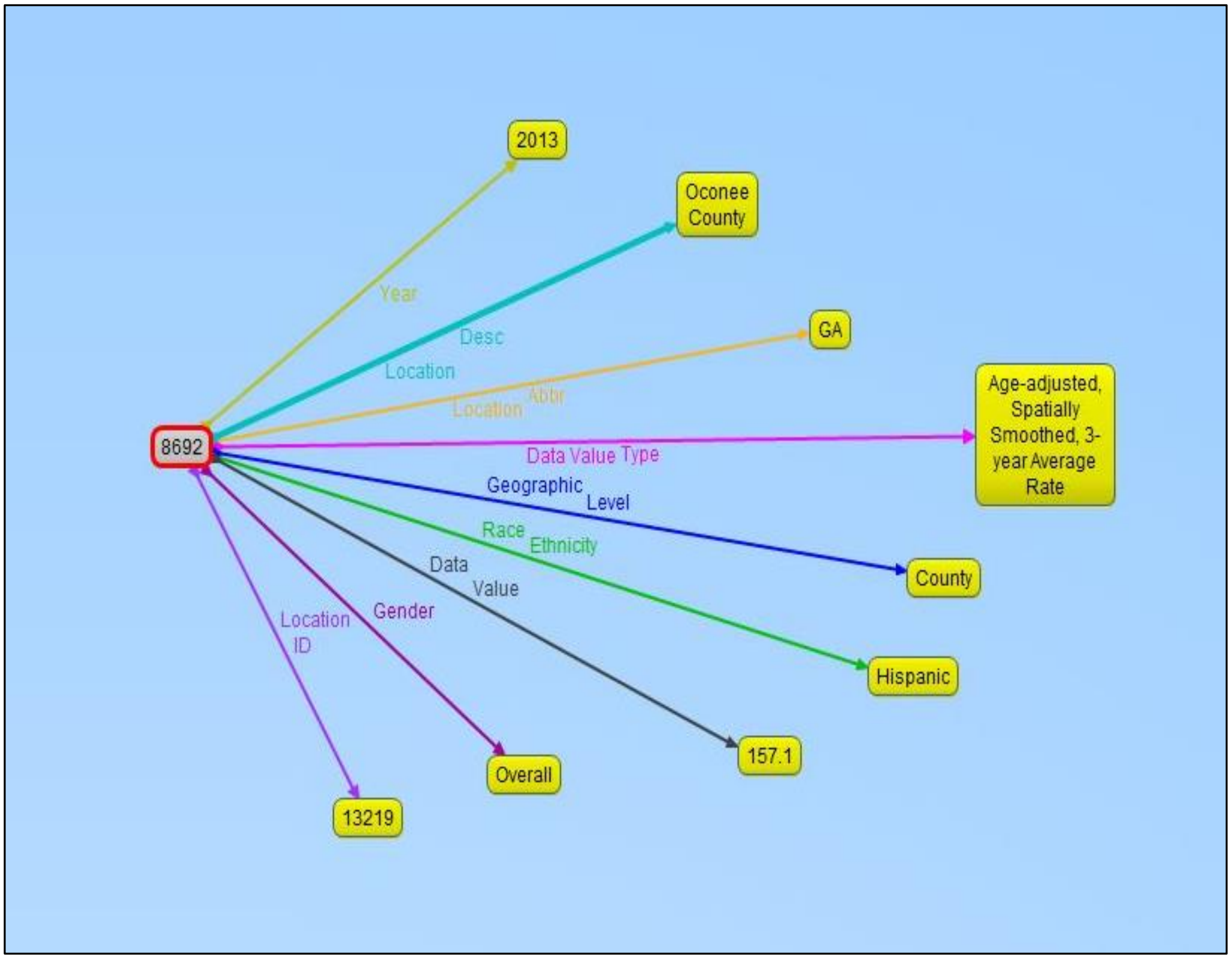


*Figure b: Heart Disease Mortality Rate dataset ontology represented using GRUFF*

## 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 Mortality rate and Average Tobacco use values for each state based on the ethnicity. For heart mortality rate dataset, we set

  - ✓ GeographicLevel property as “State”
  - ✓ Gender property as “Overall”
  - ✓ Race_Ethnicity property based on the respective race
  - ✓ And got back State’s name and its corresponding Data Value

For average tobacco use dataset, we set

- ✓ YEAR property as “2014-2015”
- ✓ Gender property as “Overall”
- ✓ Age property as “All Ages”
- ✓ Race property based on the respective race
- ✓ And got back State’s name and its corresponding Data Value

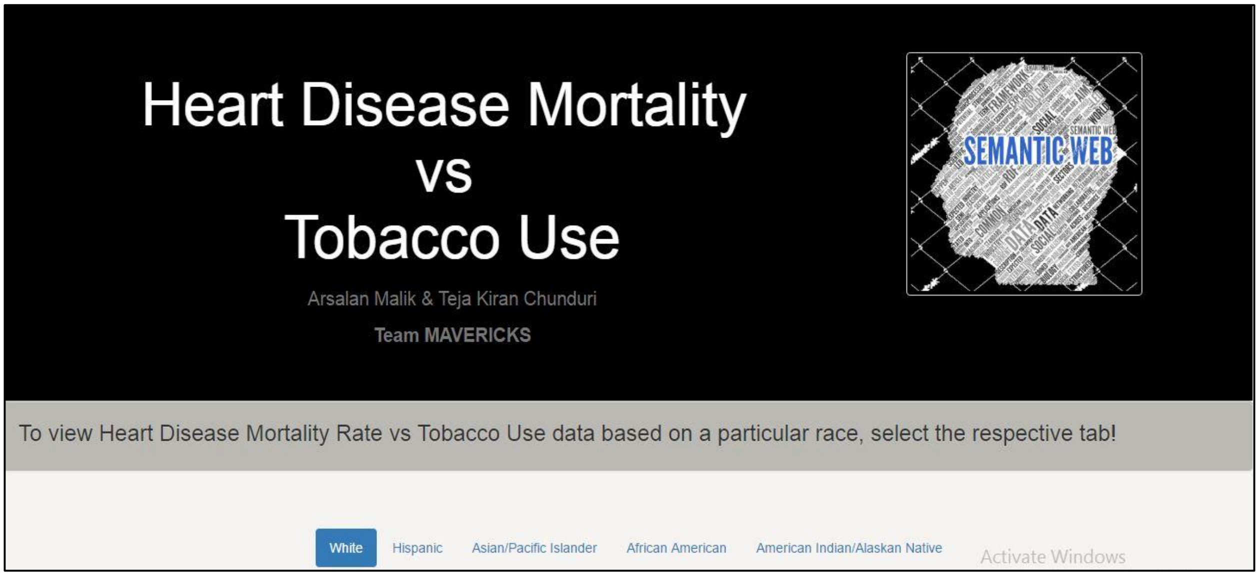
- We normalized both the datasets to a common scale by observing the data distribution and dividing each data value by the maximum value of that dataset and then multiplied by 100. As scaled data gives much better visualization.

```
SELECT (xsd:string(?state))
((((xsd:float(?datavalue1))/(xsd:float(439.4))) * xsd:float(100)) AS ?HeartMortalityRate)
((((xsd:float(?datavalue2))/(xsd:float(26))) * xsd:float(100)) AS ?AverageTobaccoUse)
```

- 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. DATA PRODUCT 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 relationship between heart disease mortality rate and average tobacco use especially for ethnic groups ‘Whites’ and ‘American Indians/Alaskan Natives’. On the other hand, for ethnic groups like ‘Asians’ and ‘African Americans’, there is a weak relationship.



Heart Disease Mortality
vs
Tobacco Use

Arsalan Malik & Teja Kiran Chunduri

Team MAVERICKS

To view Heart Disease Mortality Rate vs Tobacco Use data based on a particular race, select the respective tab!

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

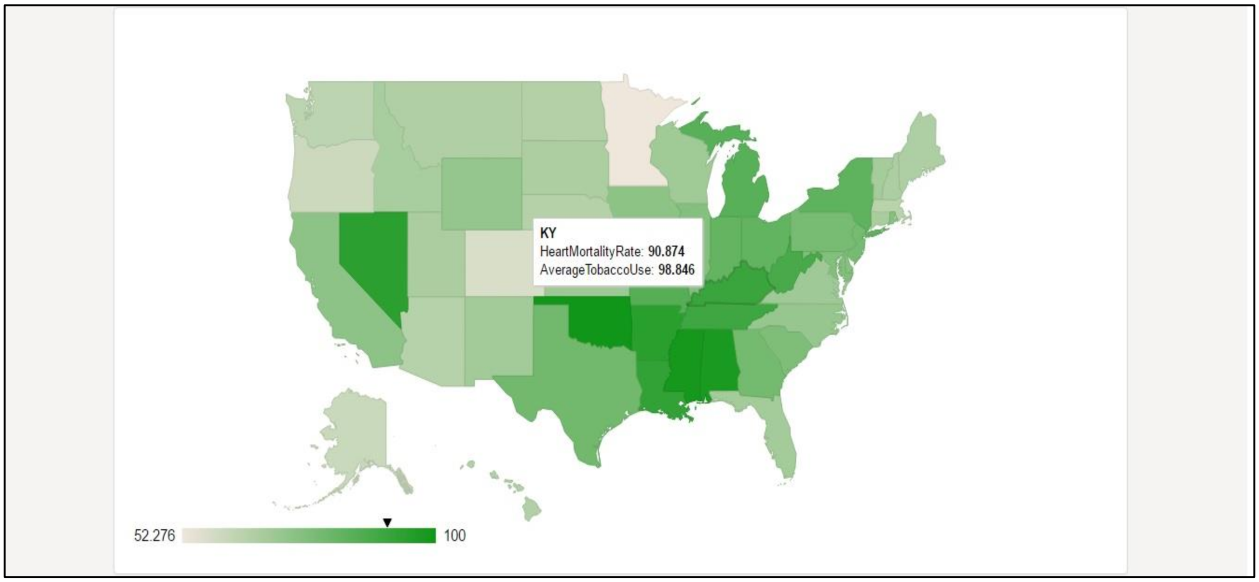


*Figure d: GeoMap representation of Heart Disease Mortality Rate and Average Tobacco Use. User can see these statistics by simply hovering over the respective state.*

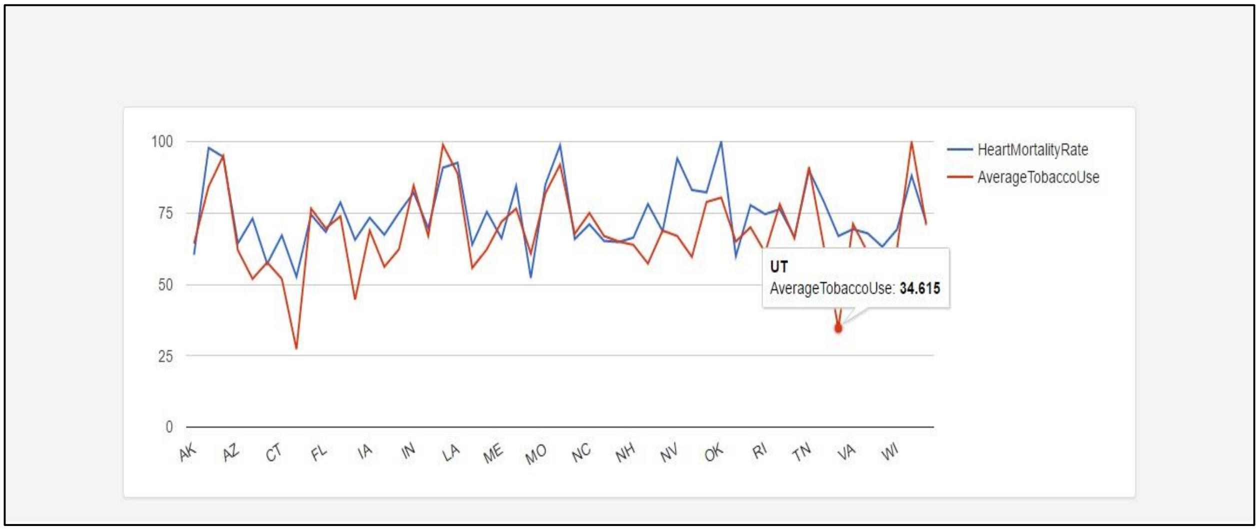


*Figure e: Line Chart representation of Heart Disease Mortality Rate and Average Tobacco Use. User can see these statistics by simply hovering over the respective state.*

# 6. SUMMARY

The goal of this project is to use Semantic Web Technologies to visualize the harmful effects the use of Tobacco has on a person's cardiovascular health. The resulting visualizations clearly shows that tobacco use has direct impact on the health of the heart and states having the highest heart mortality rate can also be seen.

## 6.1. 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.
- The conversion of CSV files into RDF files. We tried almost all the available RDF converters to accomplish this task and by using the converter RDF123, we were able to achieve our goal.