Final Project Report

Title: Study of Tobacco Consumption Trends and Effects

Team Name: ComSem's

Team Member:

Charu Arora

Josina Joy

Joydeep Roy

Manalee Panda

Type of Project: Custom Project

1. Introduction

The project aims at analyzing and understanding the data available in the following rdf datasets:

- https://catalog.data.gov/dataset/youth-tobacco-survey-yts-data
- https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi-e50c9
- http://www.americashealthrankings.org/explore/2015-annual-report

The project uses Gruff to visualize the datasets and SPARQL queries have been used to analyze the impact of tobacco in the form of cigarette smoking prevalence, cigarette smoking frequency, smokeless tobacco products use and their impact on health and chronic diseases.

2. Target Audience

Data Analysis of this project is useful to

i. Department of Health and Human Services (HHS)
The U.S. Department of Health and Human Services(HHS) protects the health of all Americans and provides essential human services. The results from our project can help the HHS to arrive at conclusions regarding how tobacco use impacts the overall health ranking of various states. It will also help the HHS to create targeted campaigns aimed at the population groups whose health is being adversely impacted by the consumption of tobacco products.

ii. General population
The results from our project showcase the adverse effects of Tobacco consumption on Health to the General Population.

3. Description of Data Sources

The datasets used for this project are:

i. Dataset 1 - Youth Tobacco Survey:

- It gives comprehensive data on both middle school and high school students regarding tobacco use, exposure to environmental tobacco smoke, smoking cessation
- Some important properties defined in dataset are Gender, Education, Data-Value, Topic, LocationDesc

ii. Dataset 2 – US Chronic Disease Indicators:

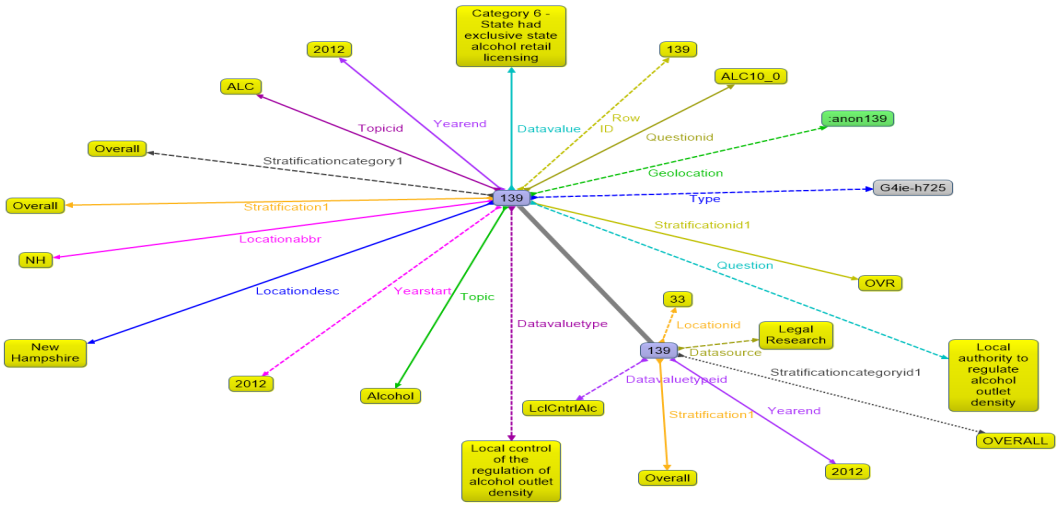
- It provides indicators like tobacco use that were developed by consensus and that allows states and territories and large metropolitan areas to uniformly define, collect, and report chronic disease data that are important to public health practice and available for states, territories and large metropolitan areas.
- Properties defined in dataset are Questions, LocationDesc, TopicDesc, Data Value, Data Value Type

iii. Dataset 3 – US Chronic Disease Indicators:

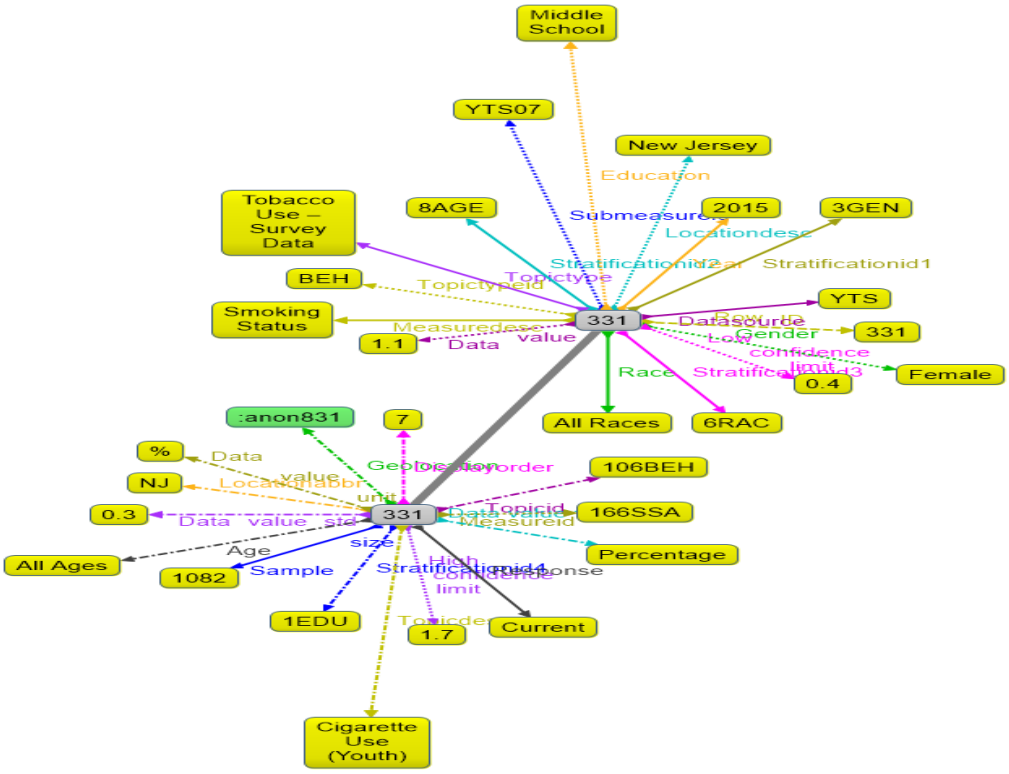
- It provides the overall health ranking of all the states in the US
- Properties defined in dataset are State_Name and Rank

4. Data Integration

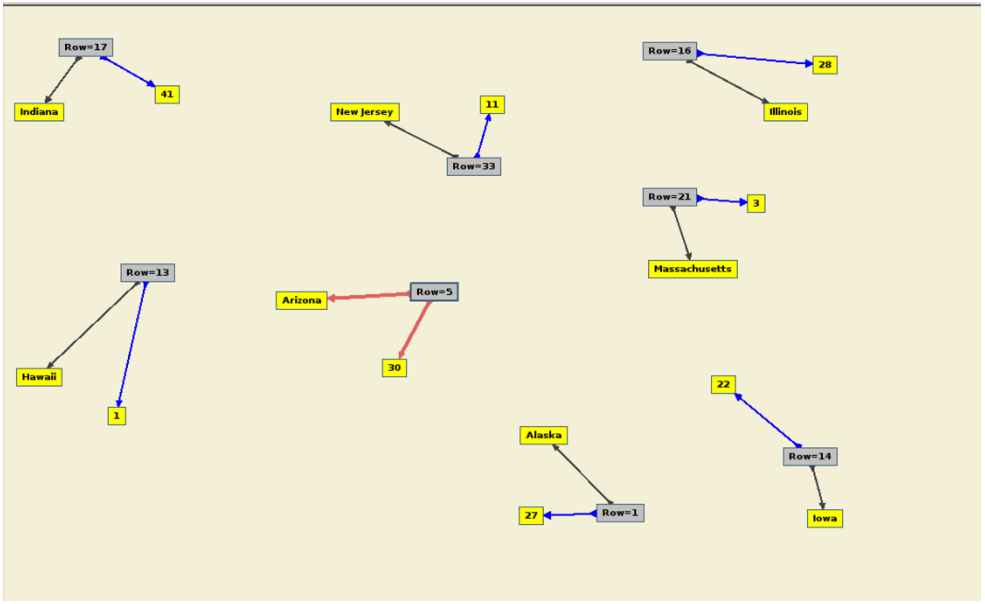
4.1 The datasets were visualized and analyzed using Gruff as follows :



Graph visualization of Chronic Disease dataset



Graph visualization of Youth Tobacco Survey dataset

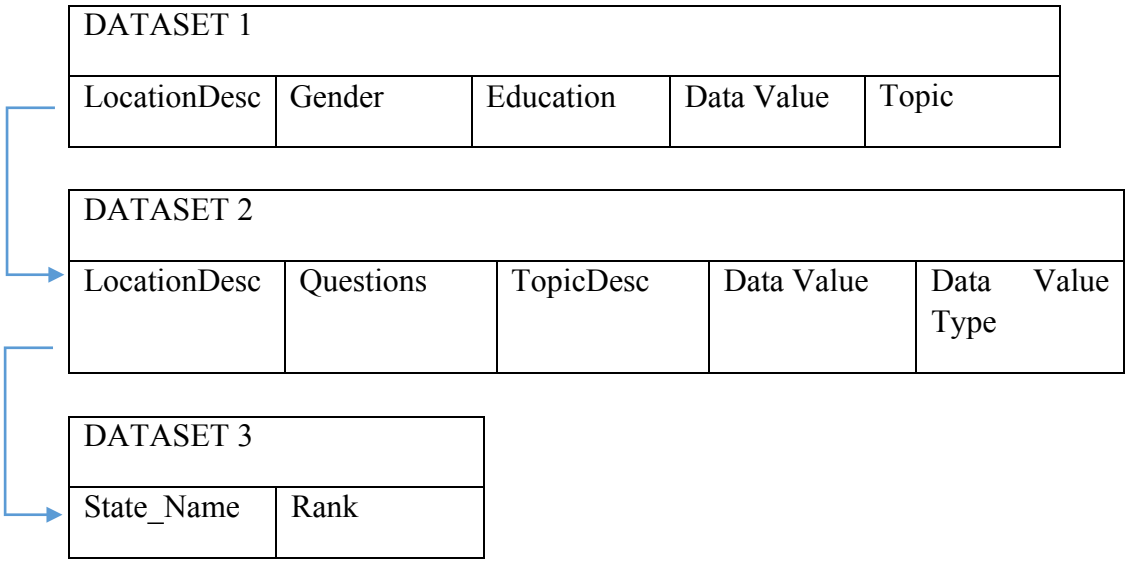


Graph visualization of Health Rankings dataset

4.2 Steps in data integration:

i. First we loaded the 3 datasets in apache-jena-fuseki server
ii. SPARQL query was generated on Dataset 1 to depict the smoking trends in Male, Female, Middle School and High School
iii. SPARQL query was generated on Dataset 2 to analyze the smoking trends in the different states in the USA
iv. SPARQL query was generated to compare the percentage of youth consuming tobacco and the total population of tobacco consumption in 14 different states in the USA. Dataset 1 and Dataset 2 were integrated for this purpose and 'LocationDesc' was the common value which was used to integrate.
v. In comparative analysis we analyzed how the number of cigarette packages sold in each state affects the health rank of the states. Dataset 2 and Dataset 3 were integrated for this purpose and 'LocationDesc/State_Name' was the common value which was used to integrate.

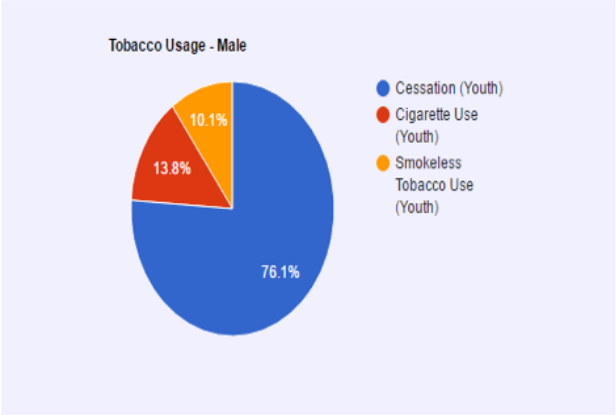
A visual representation of the values of integration:



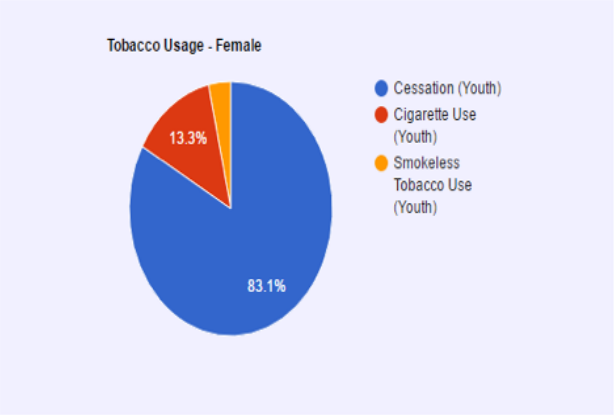
5. Data Product Results

i. From the study of tobacco consumption trends, we see that among the males in middle school and high school, 76.1% are in Cessation while 13.8 % use cigarette and 10.1% use smokeless tobacco. As compared to the males in middle school and high school, less number of females use cigarette and

smokeless tobacco and about an 83.1% are in Cessation. This result is shown in the following two graphs :



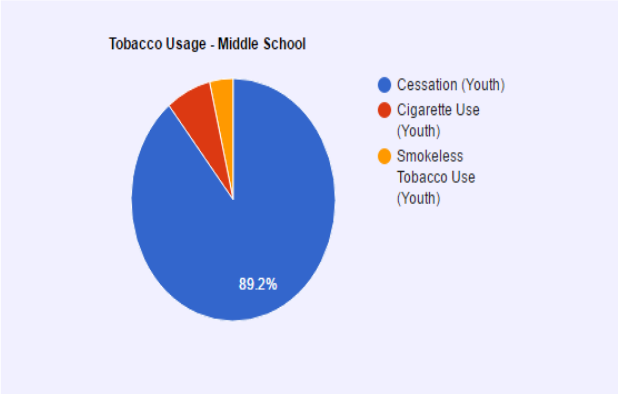
The chart shows the tobacco consumption among the males in middle school and high school living in the United States of America.



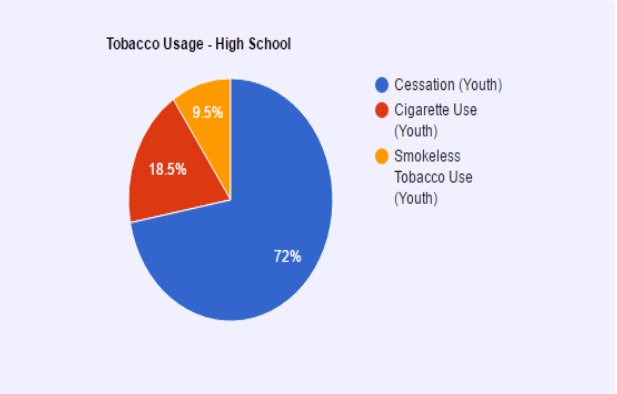
The chart shows the tobacco consumption among the females in middle school and high school living in the United States of America.

Fig. : Tobacco use -Male and Female in Middle school and High school

ii. The next comparison is in between the students in middle school and high school. From the charts below, we see that there are 89.2% of youth in Cessation in middle school while the Cessation in high school in 72%. There are more percentage of students in high school who use cigarette and smokeless tobacco than in middle school.



The chart represents the tobacco consumption among the students in middle school living in the United States of America.



The chart represents the tobacco consumption among the students in high school living in the United States of America.

Fig.: Tobacco consumption trends among students in middle school and high school

iii. The next comparison is between the tobacco consumption among youth and tobacco consumption among total population in 14 states. From the graph we see that in some states like ‘Utah’, the tobacco consumption by youth is very minimal compared to the overall consumption while in other states like ‘Georgia’, the youth make up for around one-third of the total consumption.

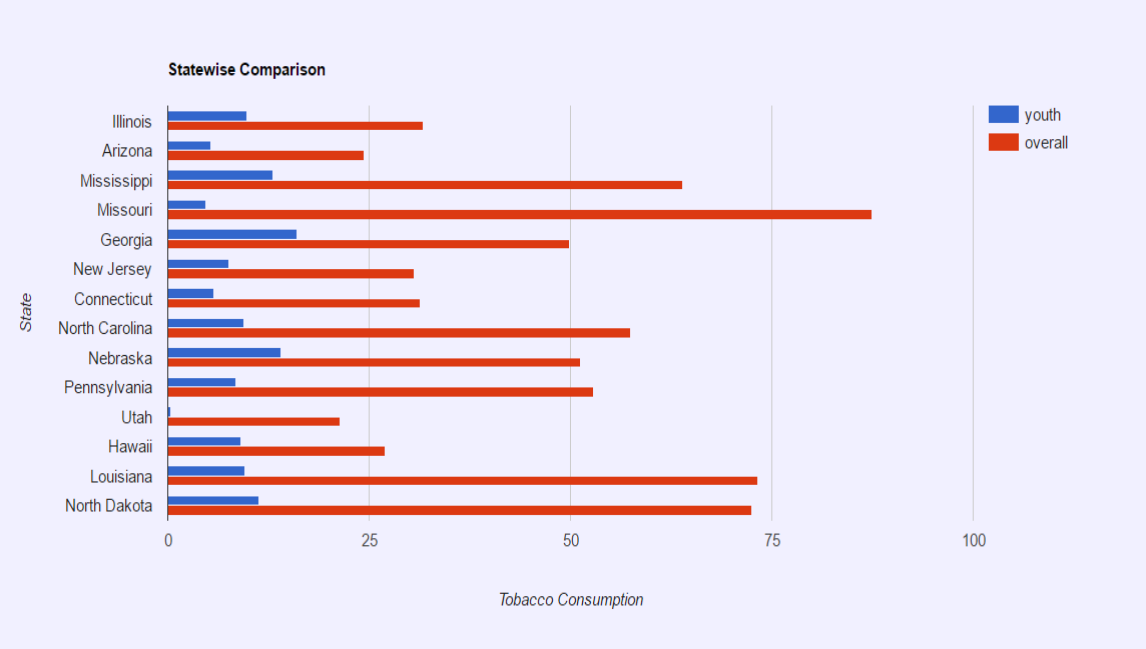
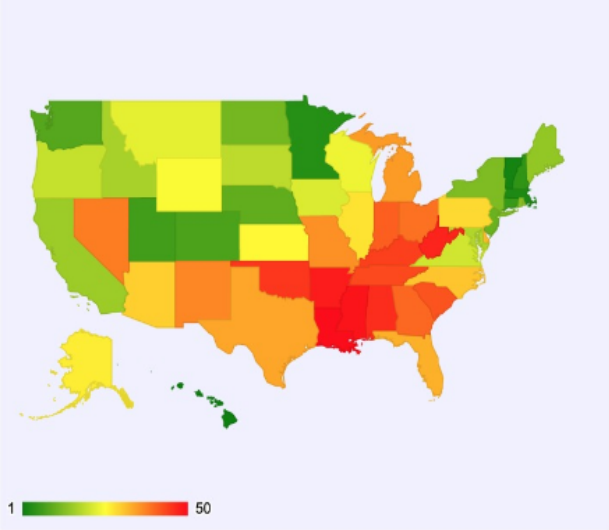
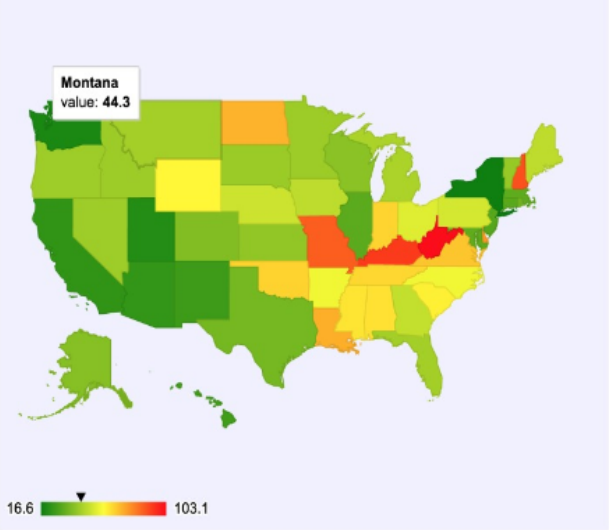


Fig.: Statewise comparison of tobacco consumption in youth and total population

iv. The next comparison is between the Health Rank in states of USA and the sale of cigarette packets among the states. From the two maps, we can conclude that the states with highest number of cigarette packets sold are the ones with poor health ranks.



The above chart shows the Health Rank of the states in USA.



The above map represents the sale of cigarette packs among the states in USA.

Fig.: Maps showing Health Rank of states and cigarette packets sold in the states

v. Also, by clicking on each state on a map, we can get a visualization of the tobacco consumption trends among various population groups in that state as shown in the following figure.

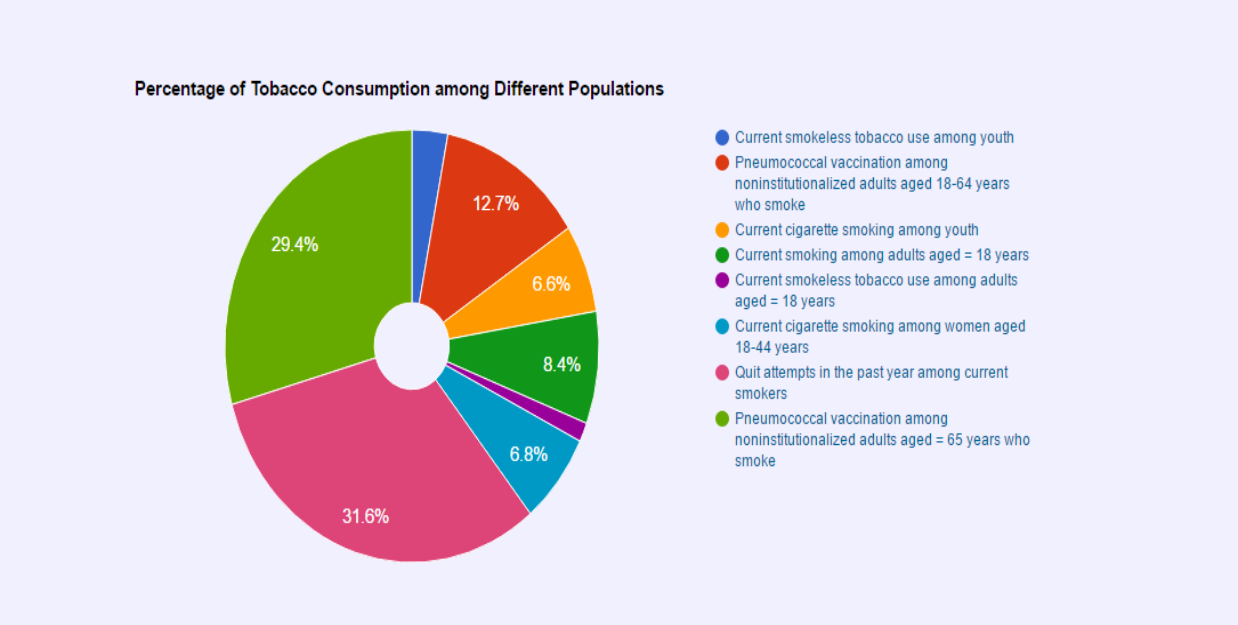


Fig.: Tobacco consumption trends in different population groups

6. Custom Project Justification

This project is different from the simple projects as we had to set up SPARQL endpoints. We used apache-jena-fuzeki server to load out datasets and set up the SPARQL endpoints.

7. Summary

In this project, we have understood and analyzed the three datasets. We visualized the data using Gruff and generated SPARQL queries on the datasets to compare the tobacco consumption trends in various population groups in the states of USA. Also, we compared the number of cigarette packets sold in every state with the Health Rank of each state. The end results of our project will be helpful to the Department of Health and Human services and also to the general population. It will make our target audience aware of the general trends in tobacco use among various population groups and also how cigarette use impacts the overall health ranking of the states. It will also help the HHS to create targeted campaigns aimed at the population groups whose health is being adversely impacted by the consumption of tobacco products.