PART I: Project Report

Title: Correlation of Fast Food Joints density and chronic diseases in the USA.

Team Information:

1. Amruta Folane (NetID: asf160130)
2. Ashish Mohapatra (NetID: axm160031)
3. Dhruv Sangvikar (NetID: dgs160230)
4. Rohit Sindhu (NetID: rks160030)

Type of Project: Custom Project

1. Introduction

The semantic web is a web of data to be shared and reused across application, enterprise, and community boundaries. This data can be linked together to develop interesting applications and generate interesting facts and information. We are using semantic web technologies like *fuseki* and gruff to mush up four datasets, query them together to generate and show results that can be of help to anyone wishing to know or analyze or compare state wise Correlation of Fast Food Joints density and chronic diseases in the USA. RDF provides the foundation for publishing and linking the data. We have taken our data as files containing triples in RDF form from the US government's data.gov website, a web portal managed and hosted by the U.S. General Services Administration, Office of Citizen Services and Innovative Technologies and have used SPARQL to query on the mashed up datasets. This gives us the result set as per the requirements marked by the user with the No. of Fast food joints in a state and other details.

2. Target Audience

This project has been developed for people who wish to know or analyze Correlation of Fast Food Joints density and chronic diseases in the USA and compare the state-wise distribution of fast foods with details like State wise patient count, State wise Claims, State wise 18+ count.

This can help users and citizens to know which states have suffered more in chronic diseases cause by Fast food joints in the state. The user can use this facility and see a direct relationship between the number of fast food joints to the number of chronic diseases, and the number of diseases to the number of claims. This will empower citizens, NGO's or local bodies with information and knowledge to check which states were most affected. With this information, they can analyze on their own the effects of the program.

3. Description of Data Sources:

The data has been taken from the data.gov, a US government's web portal managed and hosted by the U.S. General Services Administration, Office of Citizen Services and Innovative Technologies. We have taken four datasets from this portal and used SPARQL to mush them up and query on it. The datasets have been downloaded in RDF format (.rdf files) and SPARQL is used to query on these datasets.

The data sets are as follows:

i. **US Department of Health and Human Services.**

- Nutrition, Physical Activity, and Obesity - Behavioural Risk Factor Surveillance System
- Link: link
- Domain: Federal
- Description: This dataset includes data on adult's diet, physical activity, and weight status from Behavioural Risk Factor Surveillance System. This data is used for DNPAO's Data, Trends, and Maps.

ii. **US Department of Health and Human Services.**

- Specific: Centres for Disease Control and Prevention.
- Link
- Domain: Federal
- Description: This dataset includes Inpatient and Outpatient claims, Master Beneficiary Summary Files, and many other files. Indicators from this data source have been computed by personnel in CDC's Division for Heart Disease and Stroke Prevention (DHDSP). This is one of the datasets provided by the National Cardiovascular Disease Surveillance System. The system is designed to integrate multiple indicators from many data sources to provide a comprehensive picture of the public health burden of CVDs and associated risk factors in the United States.

iii. **Fast Food joint distribution in the United States.**

- Link
- Domain: Federal
- Description: Author wrote scrapers for other major fast food chains and produced a list of 50,000 locations in the US. Using the site, people can go to any location in the US and see the major fast food restaurants there. Author's intention wasn't to make it easier for people to find fast food, nor was it to criticize the ubiquity of fast food. Rather he was interested in the visualization, the graphical portrayal of information. These graphics convey something that is true about the US.

iv. **18+ Population in the United States.**

- Link
- Domain: Federal:
- Description: This dataset includes state-wise, age 18 and above population in the United States.

4. Data Integration

After getting data from completely different sources, our aim was to extract relevant real-world patterns. We have tried to do it in two ways on the 4 datasets that we have used:

1. **Relationship between 18+ population and number of fast food joints in a state**: Based on 18+ population and Fast Food location datasets, there is a clear and evident linear relationship between these 2 numbers, which we have also shown in our google

chart (Fig-2, Page-4). Clearly, the business owners are smart enough to set up more fast food joints in a place where the main target customer base is 18+ aged adults.

We achieved this using a join query on the Named Graphs created on population and fast food datasets.

2. **Relationship between the number of Fast Food joints and Obesity Rate in a state**: We have tried to run a join query between fast food and obesity datasets to extract a pattern that, if there are more fast food joints in a state, then there are more number of obese people.

   We did this using a join query on the Named Graphs created on obesity and fast food datasets.

   P.S.: We struggled to get a good dataset to do the same, but the intent of our data integration was clear.

5. Data product results

The user is provided with a link (Google Charts API) where they can see the graphs and tables to visualize density distribution of fast food joints with health diseases details along with state wise medicare claims.

SPARQL is used to query these 4 datasets present in RDF format (.rdf file). Queries for both graph and table are used and checked on Apache Jena Fuseki.

.

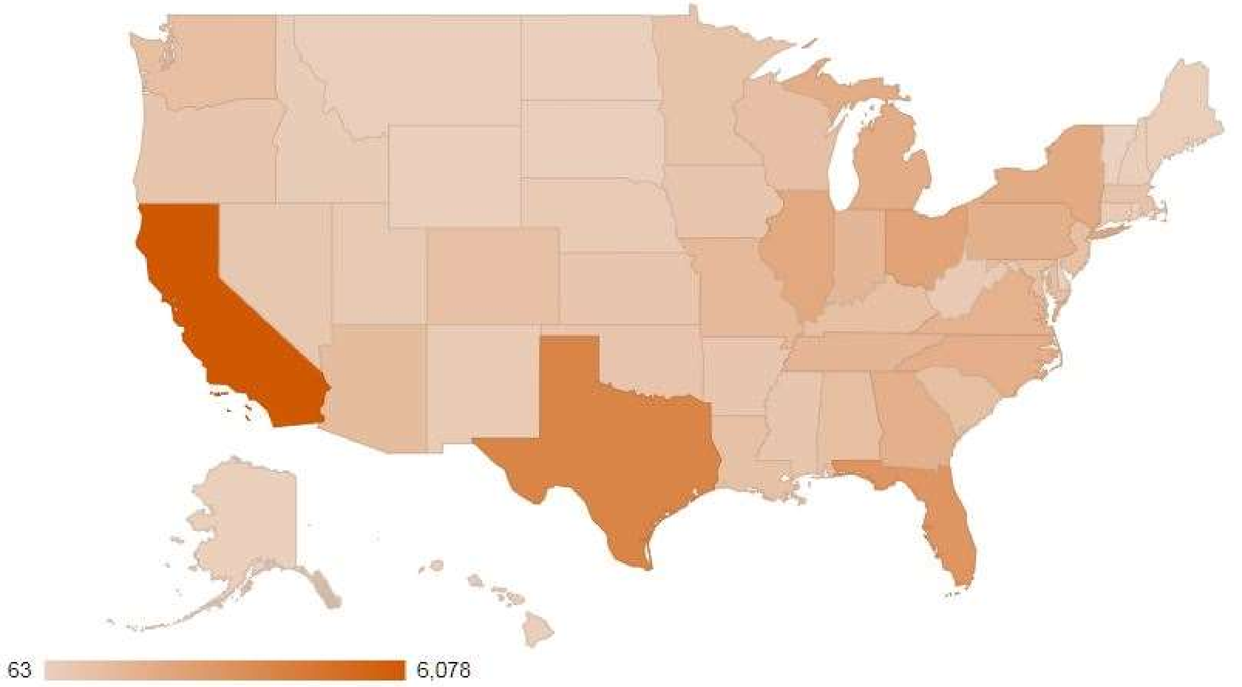

Fig 1. State-wise distribution of fast food restaurants in the USA

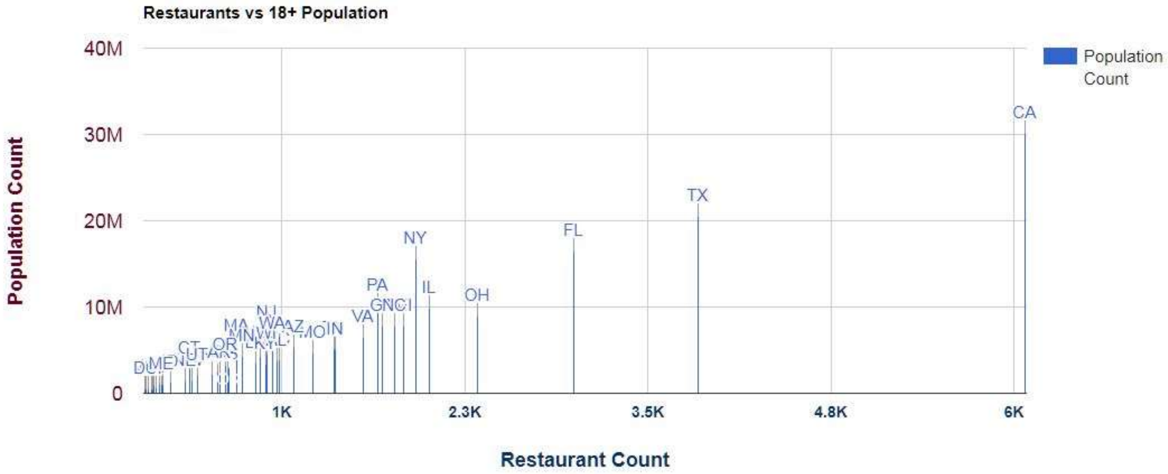

Fig 2. State-wise correlation between fast foods Restaurants and Patient count

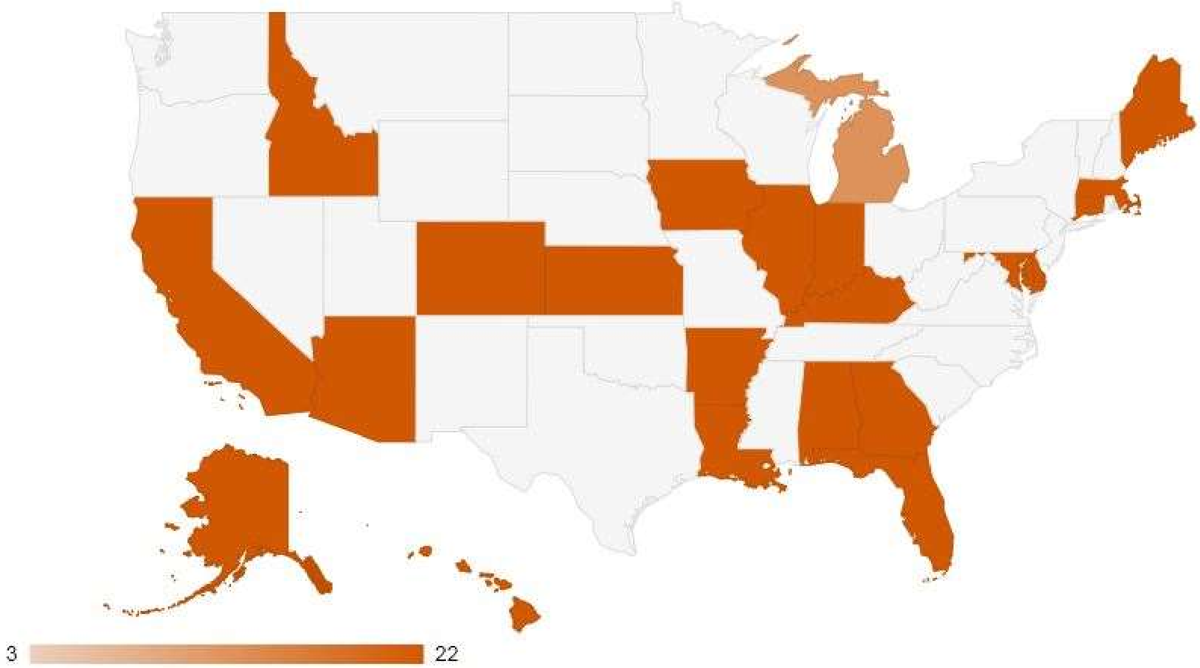

Fig 3. State-wise Patient Count.

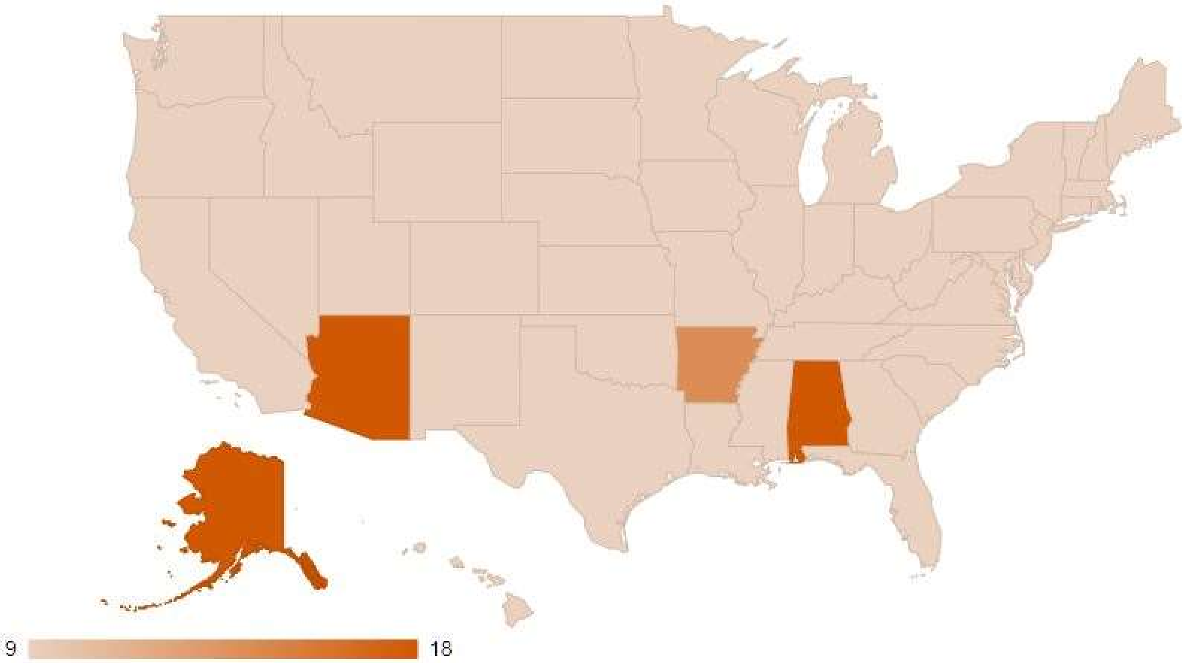

Fig 4. State-wise Medicare Claims.

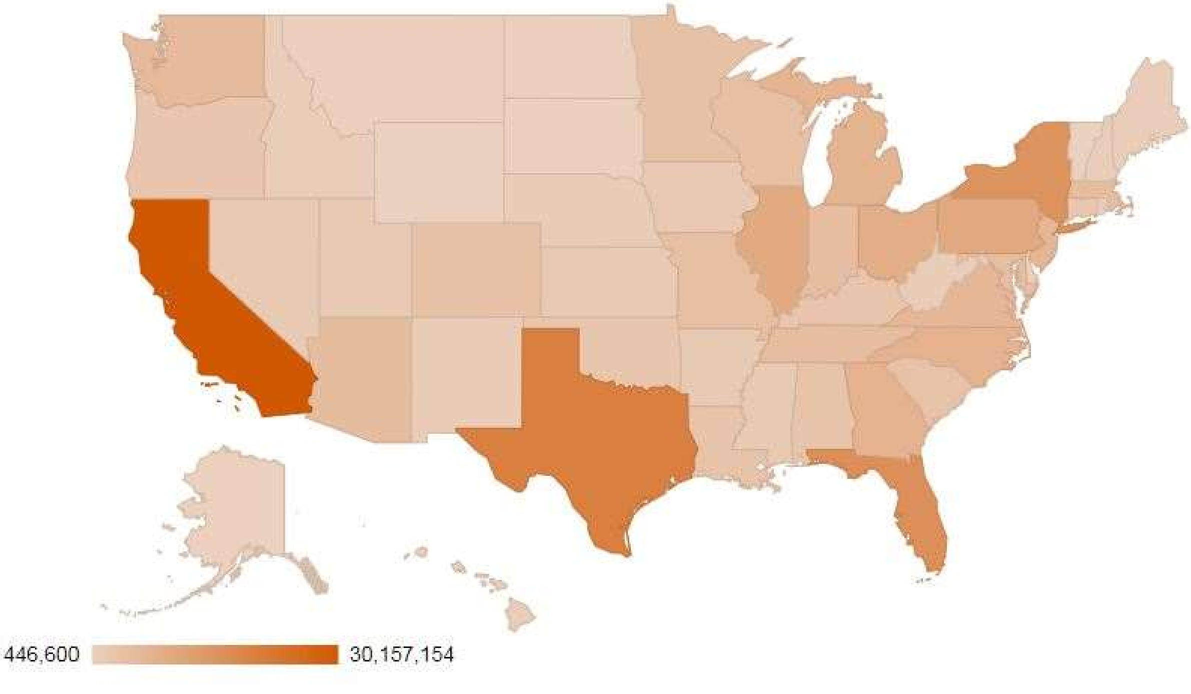

Fig 5. State-wise 18+ population

6. Custom Project Justification

With a single dataset, we couldn't derive meaningful results. Thus, had to move to a Custom project to integrate multiple datasets. We are using four datasets: Fast Food joint distribution in the United States, Centres for Disease Control and Prevention and Nutrition, Physical Activity, and Obesity - Behavioural Risk Factor Surveillance System. We have compared the fast food joints in a state with the medicare claims for that state and proposed the result.

7. Summary

This project uses four datasets taken from data.gov web portal to mash them on a common attribute. SPARQL is then used to query on these data to present a user-friendly, easy and readable visualized graph and chart. The user collects and analyzes the information he gets from these graphs and tables. This project can be of use to the reporters, common citizens gathering information, NGO's working for disease management and local bodies to gather and analyze Fast food caused chronic diseases and Medicare claims. In the future, this project can be extended to make an API where the user enters a state or year and for that year and state graphs are displayed with their information on the server thus using the extensive capabilities of semantic web technology.

PART II: Interactive Website of the Project

To view the interactive representation of the project, visit the following link*:

https://utd-abhyas.appspot.com/

*However, as our SPARQL endpoint is local and not on the server, this website works only when the endpoint is available. We will show this during the project demo.

This website demonstrates four queries that we have made on the four datasets in consideration.

Tab1: Shows the result on Query 1 as a display on the map of the USA. More the density of the color on the map, more is the number of fast food restaurants in that state. It also shows the correlation between the number of fast food restaurants in the USA and the 18+ population_count as a graph. From the graph, it can be clearly deduced that the state where the 18+ population is more, the restaurant count is more. Hence, they are directly proportional to each other.

Similarly, other tabs show the patient_count, mediclaims_count, and the 18+ population_count in that particular state. The darker the color, greater the count.

Also, the about page of the website denotes the preliminary information about the group and our project.

All these graphs are shown above in the report as figures.

Github link for the project repo: https://github.com/meets7/semanticproject

P.S. I would recommend you to check our 404 page! ;D