**FINAL REPORT**

**Team Name: Team SS**

**Team Members:** Shumei Lin (sxl157930)**,** Sowmya Bhupathiraju (sxb162130)

**Project Title: New York Violence (and Firearm) Vs Unemployment**

**Project Type: Custom Project**

**Project Topic:**

The objective of our project is to compare the violent and the unemployment statistics, and to compare the firearm and the unemployment statistic. Our goals are: 1) to find out the relationship between the number of violence and the number of unemployed people in different counties in New York State; and 2) to find out the relationship between the number of firearm and the number of unemployed people in different counties in New York State.

**1. Introduction**

We retrieve useful data from datasets containing statistics about unemployment, violent and firearm in different counties in New York State by using SPARQL Queries. We then use Google visualization to plot the retrieved data and draw conclusions based on the resulted charts. This final report contains various details of our project in different sections: the target audience of our project; description of the data sources we use; information about data integration; discussion about the results produced and finally a conclusion about the project and report.

* 1. **Resources Used**
* Gruff (For understanding the datasets)
* Rdf123 (For converting csv files to RDF)
* Apache Fuseki server (For establishing SPARQL endpoints)
* HTML (For building websites)
* JavaScript (For building websites)
* CSS (For building websites)
* Sgvizler (A Sparql Query and Google Visualization Framework)

**2. Target Audience**

* New York State Government: The New York State Government can take necessary results from our project to gain more in depth understandings about the relationship between violence and unemployment. From there, they can take necessary steps to reduce the violence rate in different counties of New York State, i.e. by providing employment.
* Social Workers: Social workers who are striving for the wellbeing of the people can have insights about the relationship between violence and unemployment.
* Police Department: The visualization helps the police department to see in which counties the violence rate are higher and also find out solutions to reduce the violent rate.

**3. Description of Data Sources**

Tools Used: Rdf123 (for converting CSV files to RDF), GRUFF (for understanding the datasets).

**3.1. Dataset-1:** Local Area Unemployment Statistics: Beginning 1976

* Source: <https://catalog.data.gov/dataset/local-area-unemployment-statistics-beginning-1976>
* Description: The Local Area Unemployment Statistics program estimates labor force statistics (labor force, employed, unemployment, unemployment rate) for New York State civilian labor force aged 16 and up.
* Format: A RDF format. The RDF file is about 23.6 MB in size.
* Number of Attributes: There are total seven attributes of the dataset.
* Number of Triples: 5000.
* Rdf description of one triple

<rdf:Description rdf:about="#62009">

<ds:area>Queensbury Town</ds:area>

<ds:year>2007</ds:year>

<ds:unemployed>500</ds:unemployed>

<ds:unemployment>3.3%</ds:unemployment>

<ds:employed>15600</ds:employed>

<ds:month>6</ds:month>

<ds:labour\_force>16200</ns:labour\_force>

</rdf:Description>

* Gruff representation

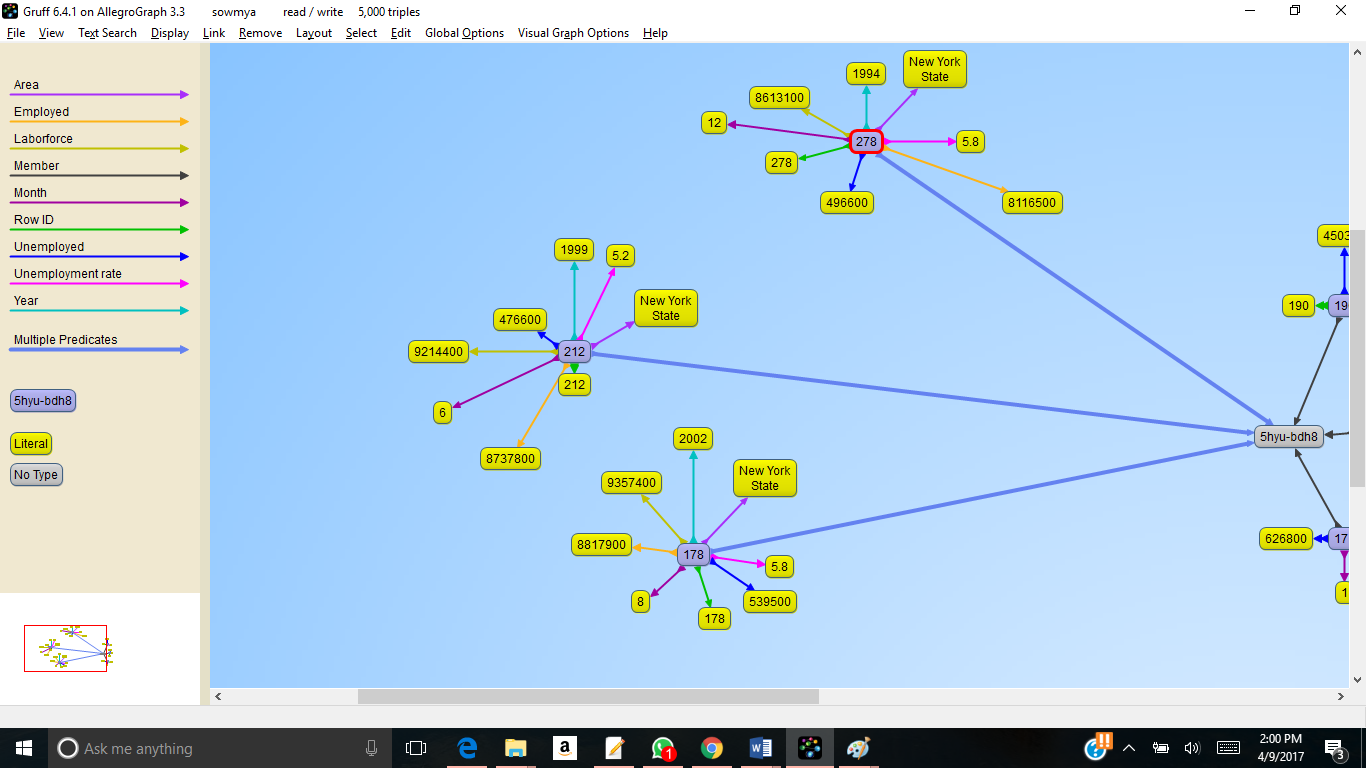


Figure 3.1: Gruff Representation of Triples of the Unemployment Dataset

**3.2. Dataset-2**: Index, Violent, Property, and Firearm Rates By County: Beginning 1990

* Source:<https://catalog.data.gov/dataset/index-violent-property-and-firearm-rates-by-county-beginning-1990>
* Description: This dataset include the violent crimes of murder/non-negligent manslaughter, forcible rape, robbery, and aggravated assault; and the property crimes of burglary, larceny, and motor vehicle theft and so on in New York State.
* Format: A RDF format. The RDF file is about 854 KB in size.
* Number of Entities: There are total eleven attributes of the dataset.
* Number of Triples: 6972.
* RDF Description of one triple

<rdf:Description rdf:about="#221">

<ns:violent\_rate>217.9</ns:violent\_rate>

<ns:index\_count>1228</ns:index\_count>

<ns:index\_rate>2388.6</ns:index\_rate>

<ns:property\_rate>2170.7</ns:property\_rate>

<ns:population>51411</ns:population>

<ns:firearm\_rate>0.0</ns:firearm\_rate>

<ns:firearm\_count>0</ns:firearm\_count>

<ns:year>2003</ns:year>

<ns:violent\_count>112</ns:violent\_count>

<ns:property\_count>1116</ns:property\_count>

<ns:county>Chenango</ns:county>

</rdf:Description>

* Gruff Representation

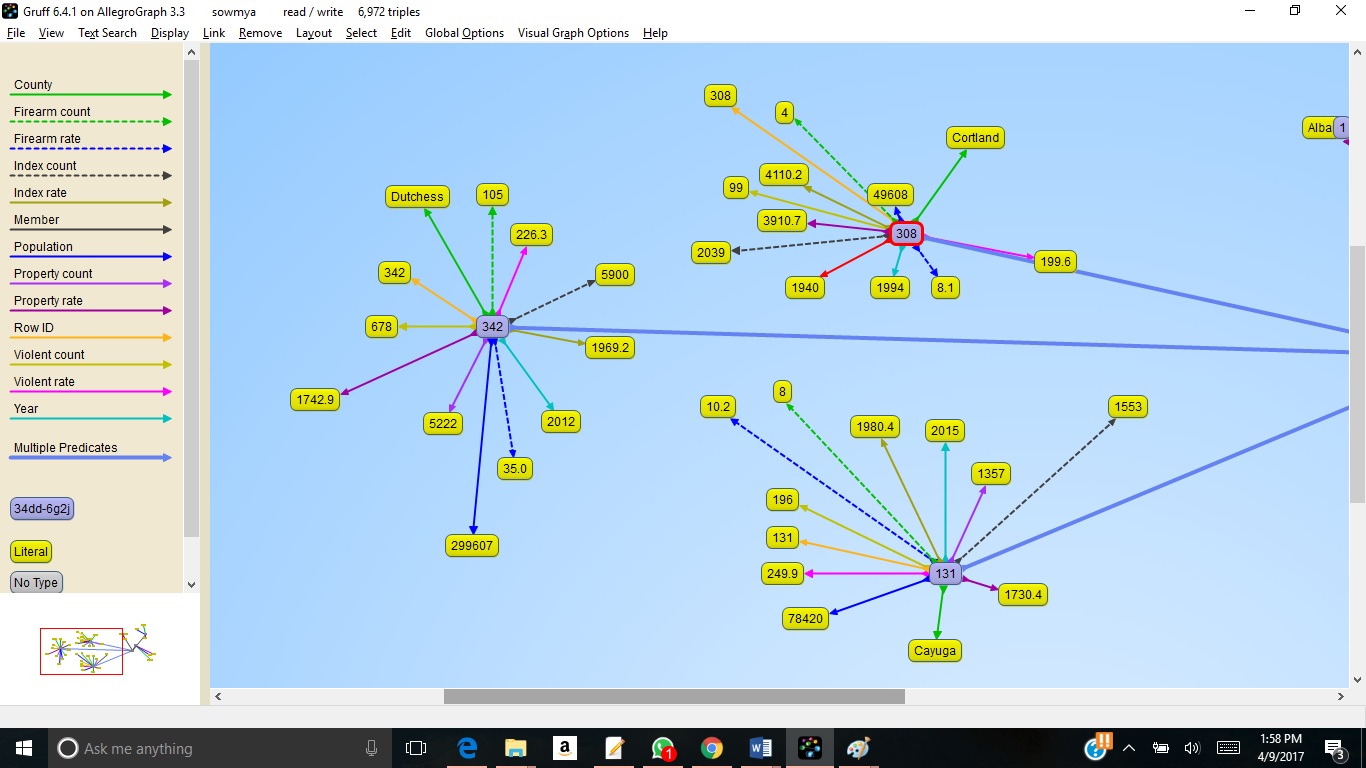


Figure 3.2: Gruff Representation of Triples of the Violence Dataset

**4. Data Integration**

The two datasets are integrated and mashed up using the common attribute “County”.

We have done two integrations:

* Comparing the unemployment count and the firearms count.
* Comparing the unemployment count and the violence count.

Tools Used: Apache Fuseiki Server, Sgvizler (A Framework for Sparql Query and Google visualization).

**4.1. First Integration**

We compared the unemployment count and the firearm count in different counties of New York State based on the common attribute “County”. The following query is used to retrieve necessary results. This is done for three years 2015, 2014 and 2013.

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?year ?county ((xsd:decimal(?firearm\_count)) AS ?firearm\_Count) (SUM((xsd:decimal(?unemployed))) AS ?Unemployment\_Count)

WHERE {

GRAPH <http://localhost:3030/project/data/violent> {

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/county> ?county.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/firearm\_count> firearm \_count.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/year> “2015”.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/year> ?year.

}

GRAPH <http://localhost:3030/project/data/unemployment> {

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/area> ?county2.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/unemployed> ?unemployed.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/year> “2015”.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/year> ?year.

FILTER regex(?county2, “County”, “i”).

}

FILTER contains(?county2, ?county)

}

GROUP BY ?year ?county ?firearm\_count

**4.2. Second Integration**

We compared the unemployment count and the violence count in different counties of New York State based on the common attribute “County”. The following query is used to retrieve necessary results. This is done for three years 2015, 2014 and 2013.

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?year ?county ((xsd:decimal(?violent\_count)) AS ?Violent\_Count) (SUM((xsd:decimal(?unemployed))) AS ?Unemployment\_Count)

WHERE {

GRAPH <http://localhost:3030/project/data/violent> {

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/county> ?county.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/violent\_count> ?violent\_count.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/year> “2015”.

?s1 <http://data.ny.gov/resource/\_34dd-6g2j/year> ?year.

}

GRAPH <http://localhost:3030/project/data/unemployment> {

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/area> ?county2.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/unemployed> ?unemployed.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/year> “2015”.

?s2 <http://data.ny.gov/resource/\_5hyu-bdh8/year> ?year.

FILTER regex(?county2, “County”, “i”).

}

FILTER contains(?county2, ?county)

}

GROUP BY ?year ?county ?violent\_count

**5.Data Product Results**

The first integration using the first SPARQL query gives the following JSON results. The results contain statistics of firearm count and unemployment count for all counties. The mash up is done on different years: 2013, 2014 and 2015.



Figure 5.1: JSON Output for Firearm Count Vs Unemployment Count

The second integration using the second SPARQL query gives the following JSON results. The results contain statistics of violent count and unemployment count for all counties. The mash up is done on different years: 2013, 2014 and 2015.



Figure 5.2: JSON Output for Violent Count Vs Unemployment Count

**5.1. Building visualization for the above data**



Figure 5.3: Google Visualization in Tabular Form for Firearm Count Vs Unemployment Count

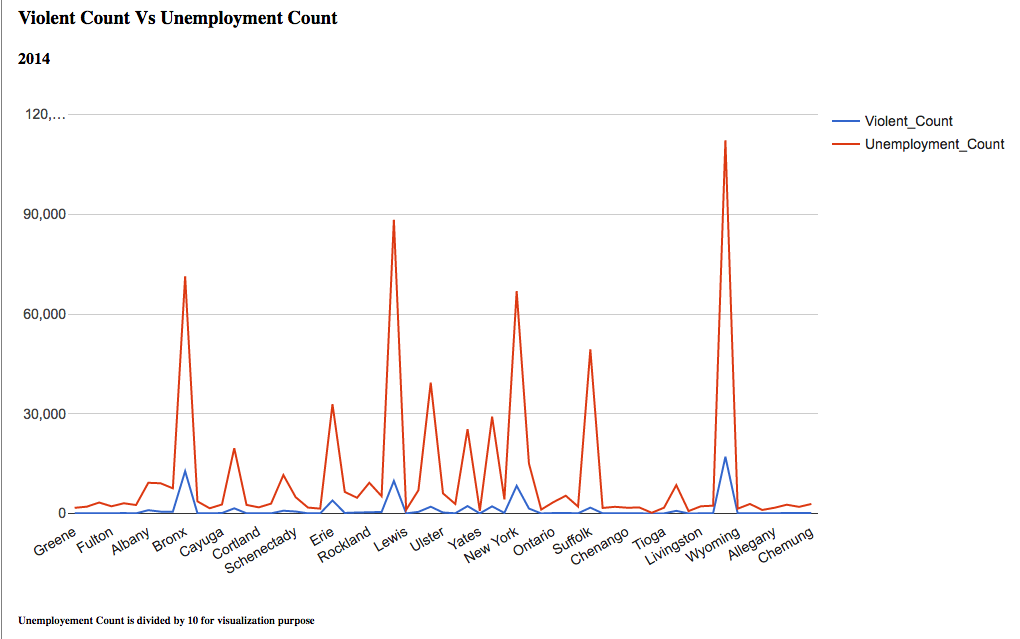


Figure 5.4: Google Visualization in Line Chart for Violent Count Vs Unemployment Count

Observation: Figure 5.4 above shows clearly that the counties have larger number of unemployment count also have larger number of violent count. From this figure we can conclude that the number of unemployment and number of violence affect each other.

Note that the unemployment count in figure 5.4 is divided by 10 in order to get a better visualization because the original unemployment count is much larger than the violent count.

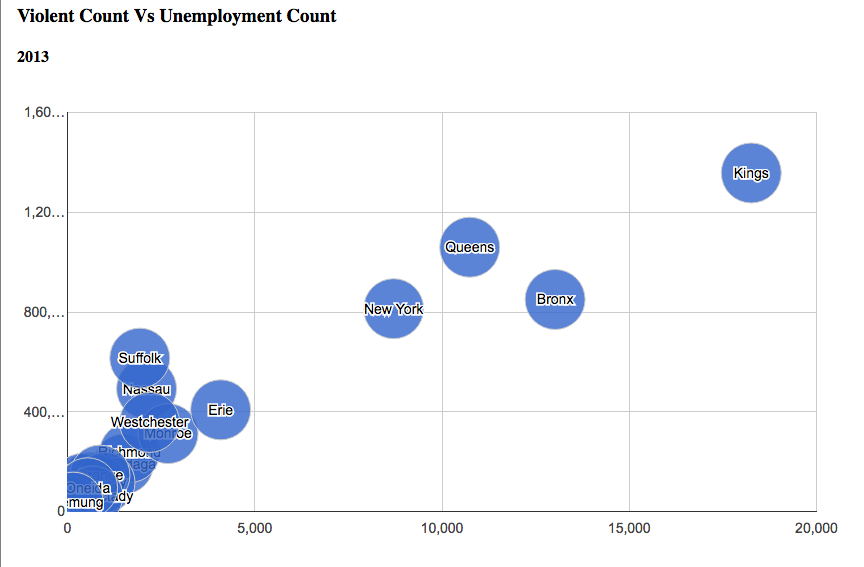


Figure 5.5: Google Visualization in Bubble Chart for Violent Count Vs Unemployment Count

Observation: We can acquire the same observation from figure 5.5 above -- the counties have larger number of unemployment count also have larger number of violent count. From this figure we can conclude that the number of unemployment and number of violence affect each other.

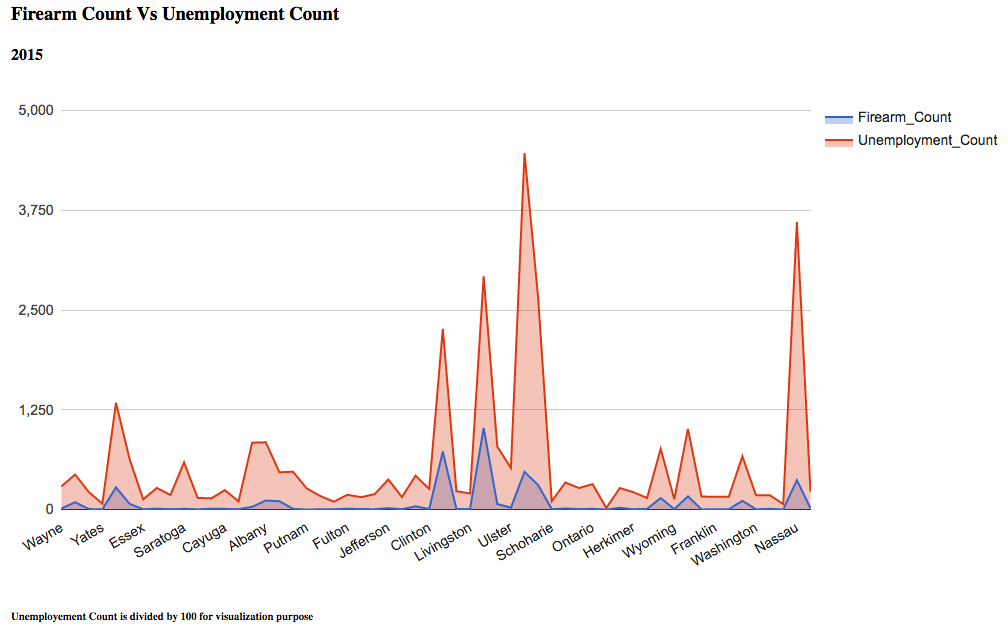


Figure 5.6: Google Visualization in Area Chart for Firearm Count Vs Unemployment Count

Observation: The relationship between firearm count and unemployment count is the same as the relationship between violent count and unemployment count. From figure 5.6 above, we can observe that counties with higher number of unemployment have higher number of firearm count.

Note that the unemployment count in figure 5.6 is divided by 100 in order to get a better visualization because the original unemployment count is much larger than the firearm count.

**6. Custom Project Justification**

The justifications to say that this is a custom project are:

• We established our own SPARQL endpoints for our datasets with Apache Jena FUSEKI server. The SPARQL endpoints are hosted in local host, and can be accessed at http://localhost:3030. This is the first major difference between our project and a simple project because we built our own SPARQL endpoints instead of using the available LOGD endpoints.

• The datasets in CVS format are downloaded from Data.gov. Then the datasets are converted from CSV format to RDF format using Rdf123. The converted datasets are then uploaded to the FUSEKI server. This is the second major difference between our project and a simple project because we need to convert our datasets and then upload it manually to the server; while for a simple project, the datasets are already available to use on the LOGD site and the LOGD SPARQL endpoints.

• The Google visualization uses JSON output from our SPARQL queries on the datasets hosed on the FUSEKI endpoints on local host, whereas in a simple (LOGD) project the query is present on the LOGD site.

**7. Summary**

**7.1 Flow of Project**

* Initially, research was conducted on searching datasets to be used in the project in order to get meaningful and useful information and results. We have looked through various datasets in Data.gov website and finally chose the datasets mentioned in section 3. Initially we started as a simple project but later we converted it to a custom project because there were no SPARQL endpoints for our chosen datasets.
* Secondly, a detailed study on the RDF form of our datasets is done. We used gruff to get a visual understanding of the semantics of our datasets. We faced a few challenges here while converting the CVS files to RDF files. RDF123 is used for converting.
* Then a study is conducted on how to integrate the datasets in order to get meaningful and useful information. The Apache Jena Fuseki server is used to host the SPARQL endpoints for the datasets. We wrote some SPARQL queries to query interesting and useful data. Then JSON results are obtained from the queries.
* Finally, we used Sgvizler (a framework that utilize SPARQL queries with Google Visualization API) to show visualization of the results. A webpage is created using html, css and JavaScript to show our visualization results.

**7.2. Challenges Faced**

* The first challenge faced was to select the datasets in order to show useful and meaningful information. We have searched various datasets, and we compared and integrated them in different ways in order to see which datasets are best for our project.
* The second challenge faced was that when we started to work on the datasets we realized that the RDF files downloaded from Data.gov do not contain complete information. Therefore, we converted the CSV files downloaded from the website to RDF files using RDF123. We have tried several softwares but only Rdf123 gave descent results.
* The third challenge was to integrate two datasets using the common attribute -- “County”. Although both datasets have the “County” attribute, the counties are written in different formats in two datasets. For example, one dataset writes as “Kings” and the other dataset writes as “Kings County”. Therefore, we used “FILTER regex” and “FILTER contains” to match the counties.

**7.3 Conclusion**

The visualization results obtained are very interesting and useful and are as what we expected. The counties have higher unemployment rate also have higher violent rate and more firearm counts. The relationship between unemployment and violence, and the relationship between unemployment and firearm are clearly showed from our visualization results. The results from our project will benefit our target audience.

From this project we have learned a lot and have more understandings about semantic web. We have learned how to convert datasets to RDF format, and then build our own SPARQL endpoints with the RDF datasets. We have learned how to integrate datasets in order to get meaningful results. We are now more confident in writing SPARQL queries and have more knowledge about how to present query results using Google visualization.