# [C:\Users\jwoo5\AppData\Local\Temp\templateTermTutorial.html](http://www.calstatela.edu/centers/hipic) CIS4560-01 Term Project Tutorial https://avatars2.githubusercontent.com/u/4156894?v=3&s=100

California State University, Los Angeles

#### Authors: Luis Cardenas, Jocelyn Gallo, Raymond Ko, Diana Hoang, Eric Kha

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 12/1/2017

**Lab: Chicago’s Crime and Social Economic Data Analysis using Hadoop, Apache Pig, and Tableau**

**Objectives**

In this hands-on lab, you will:

1. Download the data
2. Upload data files
3. Load data files into Pig
4. Define relations
5. Filter data
6. Group data
7. Join relations
8. Store data to HDFS file
9. Use Tableau for visualizations

**Platform Spec**

* IBM Bluemix BigInsights
* Cluster Version: 4.2
* # of Management Nodes: 1
* # of Data Nodes: 1
* Management Node’s VCPU: 12
* Data Node’s VCPU: 4
* Management Node’s RAM: 42 GB
* Data Node’ s RAM: 24GB

Step 1: Download the data

You need to download “crime.csv” and “social-econ.csv” to compare the two data later. From BigInsights terminal, download “crime.csv” and “social-econ.csv” using the following shell commands:

$ wget -O crime.csv <https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD>

$ wget -O social-econ.csv <https://data.cityofchicago.org/api/views/kn9c-c2s2/rows.csv?accessType=DOWNLOAD>

Step 2: Upload the data files

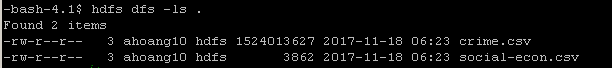
After the files are downloaded, you need to put/upload them to your user’s directory. Run the following HDFS commands to upload those csv files:

$ Hdfs dfs -put crime.csv .

$ hdfs dfs -put social-econ.csv .

Navigate to /user/your\_name to make sure if it has the files uploaded by running the following command:

$ hdfs dfs -ls .

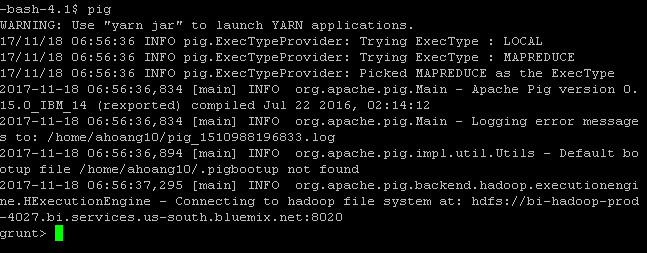


Step 3: Load data files into Pig

In this step, you will write a script to load the data files into pig and define relations.

First, open the **Pig** interface Grunt by typing in “pig” in the terminal:

$ pig



Once you’re in the pig interface, you can load the csv files. You need to load both “crime.csv” and “social-econ.csv” with a defined schema using the following scripts:

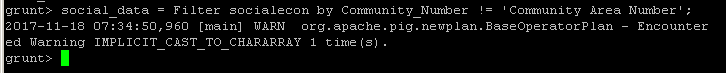
socialecon = LOAD 'social-econ.csv' USING PigStorage(',') AS (Community\_Number,COMMUNITY\_NAME,HOUSING\_CROWDED, HOUSEHOLDS\_BELOW\_POVERTY, AGE16\_UNEMPLOYED, AGE25\_WITHOUT\_HSDIPLOMA, AGE18OR64, CAPITA\_INCOME , HARDSHIP\_INDEX);

This script loads in 'social-econ.csv' and defines field 1 as Community\_Number, field 2 as COMMUNITY\_NAME, and so on. The field names can be obtained from <https://catalog.data.gov/dataset/census-data-selected-socioeconomic-indicators-in-chicago-2008-2012-36e55>.

Since this file has a header, you need to filter out the header by typing the following script in the command:

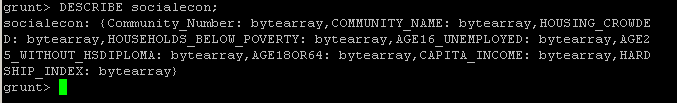
social\_data = Filter socialecon by Community\_Number != 'Community Area Number';

This script will extract out the Community\_Number field with value as 'Community Area Number', which is the first row or the header row.



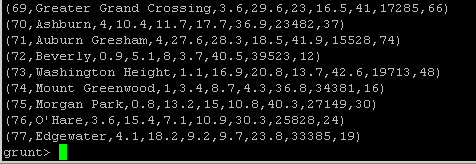
You can type in the following script to view the structure of the data that were pass to relation “socialecon”:

DESCRIBE socialecon;



To view the data inside “socialecon”, run the following script:

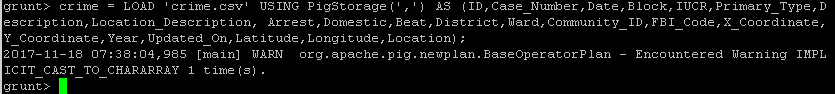
DUMP social\_data;



Now, you need to load 'crime.csv' with a schema using the following script:

crime = LOAD 'crime.csv' USING PigStorage(',') AS (ID,Case\_Number,Date,Block,IUCR,Primary\_Type,Description,Location\_Description, Arrest,Domestic,Beat,District,Ward,Community\_ID,FBI\_Code,X\_Coordinate,Y\_Coordinate,Year,Updated\_On,Latitude,Longitude,Location);

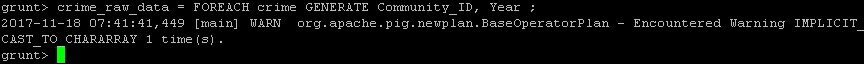
This script loads 'crime.csv' and defines field 1 as ID, field 2 as Case\_Number, and so on. The field names for ‘crime.csv’ can be obtained from <https://catalog.data.gov/dataset/crimes-2001-to-present-398a4>.



For this lab, you only need Community\_ID and Year attribute from relation “crime”. You can extract Community\_ID and Year using the following script:

crime\_raw\_data = FOREACH crime GENERATE Community\_ID, Year ;

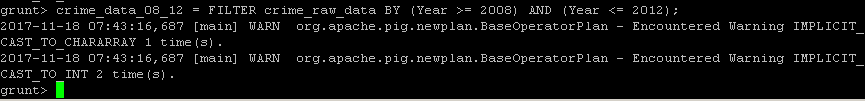
This script will keep Community\_ID and Year from the relation “crime” data and pass the result to relation “crime\_raw\_data”.



Step 4: Filter data

Since the data set for “social-econ.csv” only holds records for 2008 to 2012, you need to filter the crime data to keep only data occurred from 2008 to 2012. You can do that by running the following script:

crime\_data\_08\_12 = FILTER crime\_raw\_data BY (Year >= 2008) AND (Year <= 2012);



Since Community\_ID should only be from 1 to 77, any data that has Community\_ID's field as 0 or empty should be excluded from the final result.

First, you need to find how many results will be omitted using the following script:

omitted\_data = FILTER crime\_data\_08\_12 BY (Community\_ID is null) OR (Community\_ID == 0);

This script pulls out data with Community\_ID as null or 0 and creates a new relation “omitted\_data”.

The following script will group the data by Community\_ID, so the data can be grouped by 0 or null/empty.

grp\_omitted\_data = group omitted\_data BY Community\_ID;

Once the data is grouped, you can generate the total count for Community\_ID with field as "0" or empty using the following script:

cnt\_omitted\_data = FOREACH grp\_omitted\_data GENERATE group AS Community\_ID, COUNT(omitted\_data.Community\_ID);

You can view the total count of the omitted data by typing:

DUMP cnt\_omitted\_data;



The result shows a total of 30 counts for Community\_ID with “0” as the value and 0 count for empty value. This suggests that not many data is lost as the result of Community\_ID being marked as 0 or empty.

Finally, you can filter out data with Community\_ID's field as 0 or empty

crime\_data = FILTER crime\_data\_08\_12 BY (Community\_ID is not null) AND (Community\_ID != 0);

Step 5: Group data

To show the count for the crimes each year by Community\_ID, you first need to group crime\_data by Community\_ID and Year

grpd\_comm\_year = GROUP crime\_data BY (Community\_ID,Year);

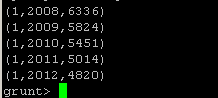
To generate a total count of crime occurred for each year from 2008 to 2012 from each Community\_ID, run the following script:

comm\_year= FOREACH grpd\_comm\_year{ year = crime\_data.Year; GENERATE group.Community\_ID AS Community\_ID, group.Year AS year, COUNT(year) AS cnt\_crime\_per\_year ; };

To display a sample set of the data that was generated, run the following script to return only 5 results from the data.

limit\_comm\_year\_crime = LIMIT comm\_year 5;

DUMP limit\_comm\_year\_crime;

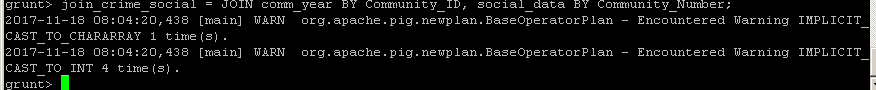


The results display the Community\_ID, year, and total crime occurred corresponding to that year.

Step 6: Join relations

To compare the crime rates and social economics among each community, you can join comm\_year and social\_data together using the following script:

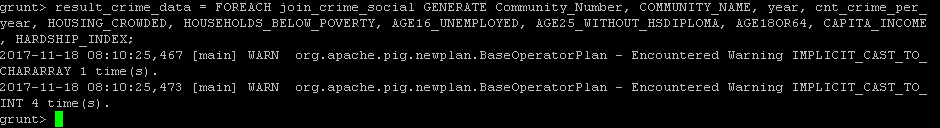
join\_crime\_social = JOIN comm\_year BY Community\_ID, social\_data BY Community\_Number;



join\_crime\_data = FOREACH join\_crime\_social GENERATE Community\_ID, year, cnt\_crime\_per\_year,Community\_Number,COMMUNITY\_NAME,HOUSING\_CROWDED,HOUSEHOLDS\_BELOW\_POVERTY, AGE16\_UNEMPLOYED, AGE25\_WITHOUT\_HSDIPLOMA, AGE18OR64, CAPITA\_INCOME, HARDSHIP\_INDEX;

Remove Community\_ID attribute from the data because it's a duplicate of Community\_Number. You can do it by defining a new relation based on the existing one:

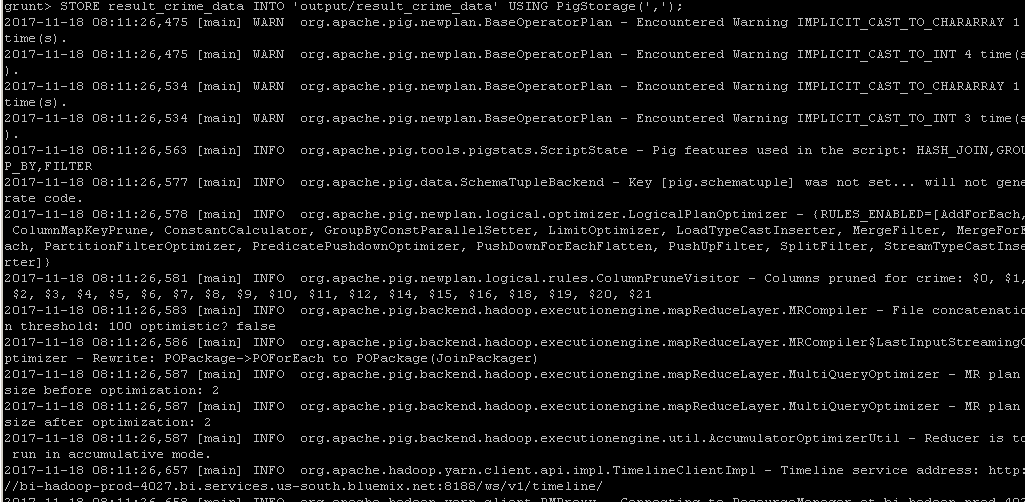
result\_crime\_data = FOREACH join\_crime\_social GENERATE Community\_Number, COMMUNITY\_NAME, year, cnt\_crime\_per\_year, HOUSING\_CROWDED, HOUSEHOLDS\_BELOW\_POVERTY, AGE16\_UNEMPLOYED, AGE25\_WITHOUT\_HSDIPLOMA, AGE18OR64, CAPITA\_INCOME, HARDSHIP\_INDEX;



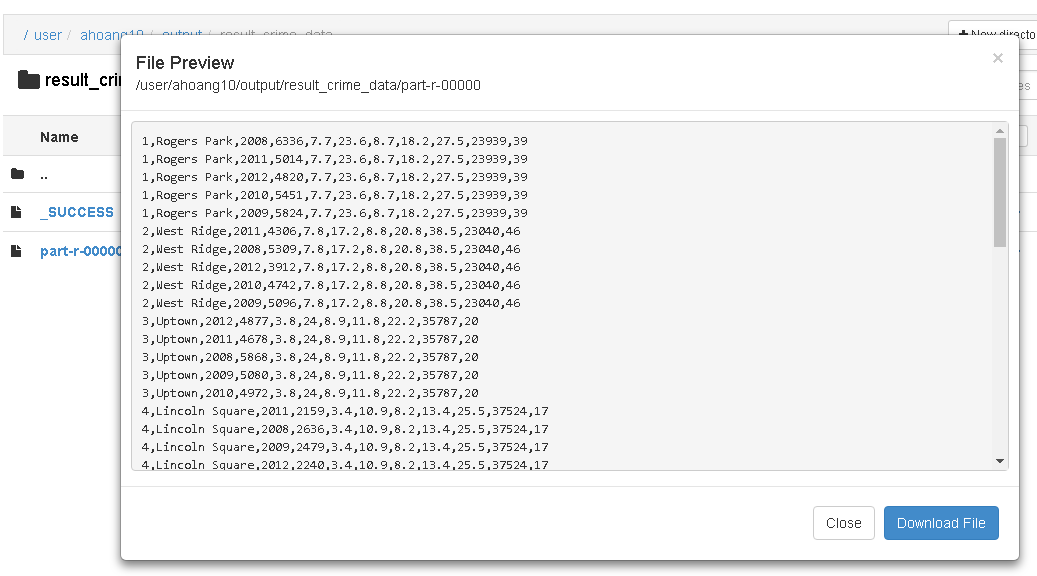
Step 7: Export joint data

Export data to output/result\_crime\_data

STORE result\_crime\_data INTO 'output/result\_crime\_data' USING PigStorage(',');



Step 8: Download output from Ambari File Browser

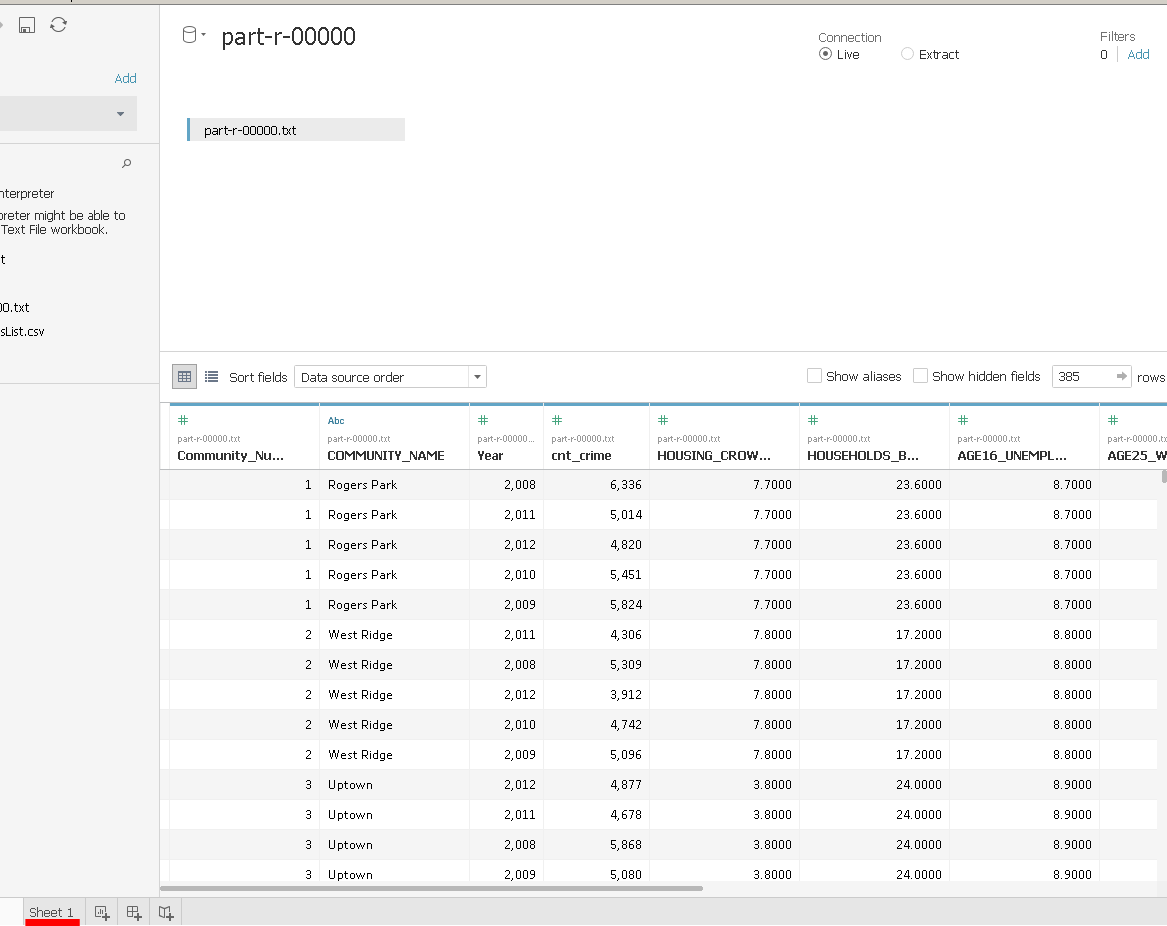


Step 9: Visualization

Once “part-r-00000” file is downloaded, open the file and save as “txt”.

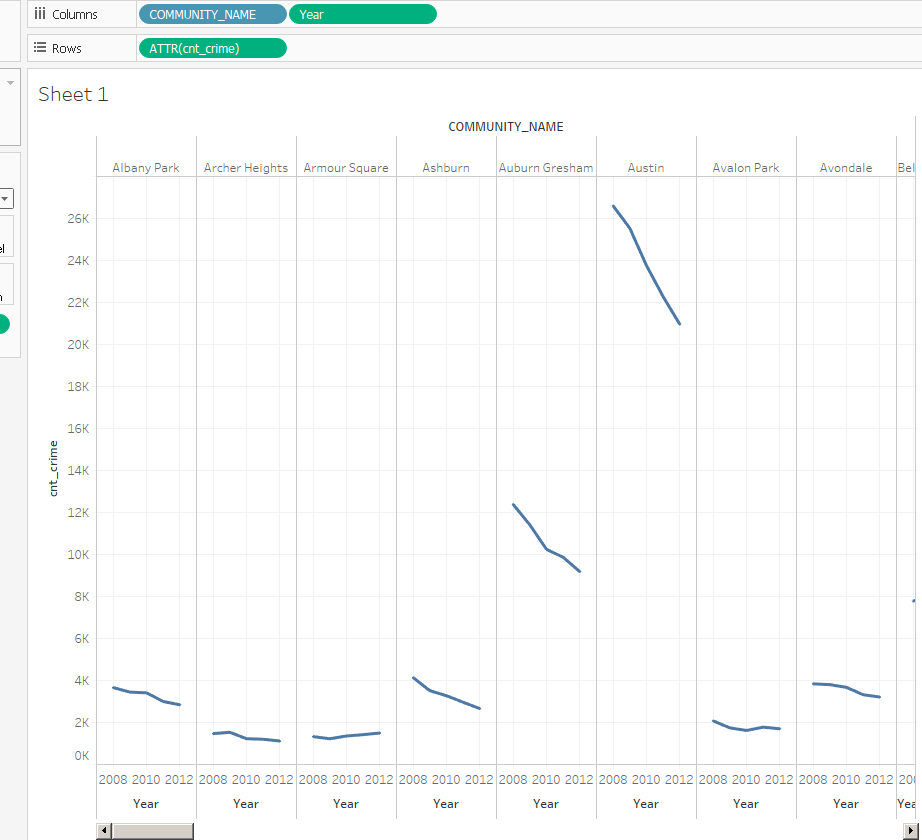
Open Tableau and open “part-r-00000.txt”.

Add/Label the fields as “Community\_Number, COMMUNITY\_NAME,year, cnt\_crime\_per\_year,HOUSING\_CROWDED,HOUSEHOLDS\_BELOW\_POVERTY,AGE16\_UNEMPLOYED,AGE25\_WITHOUT\_HSDIPLOMA,AGE18OR64,CAPITA\_INCOME,HARDSHIP\_INDEX”

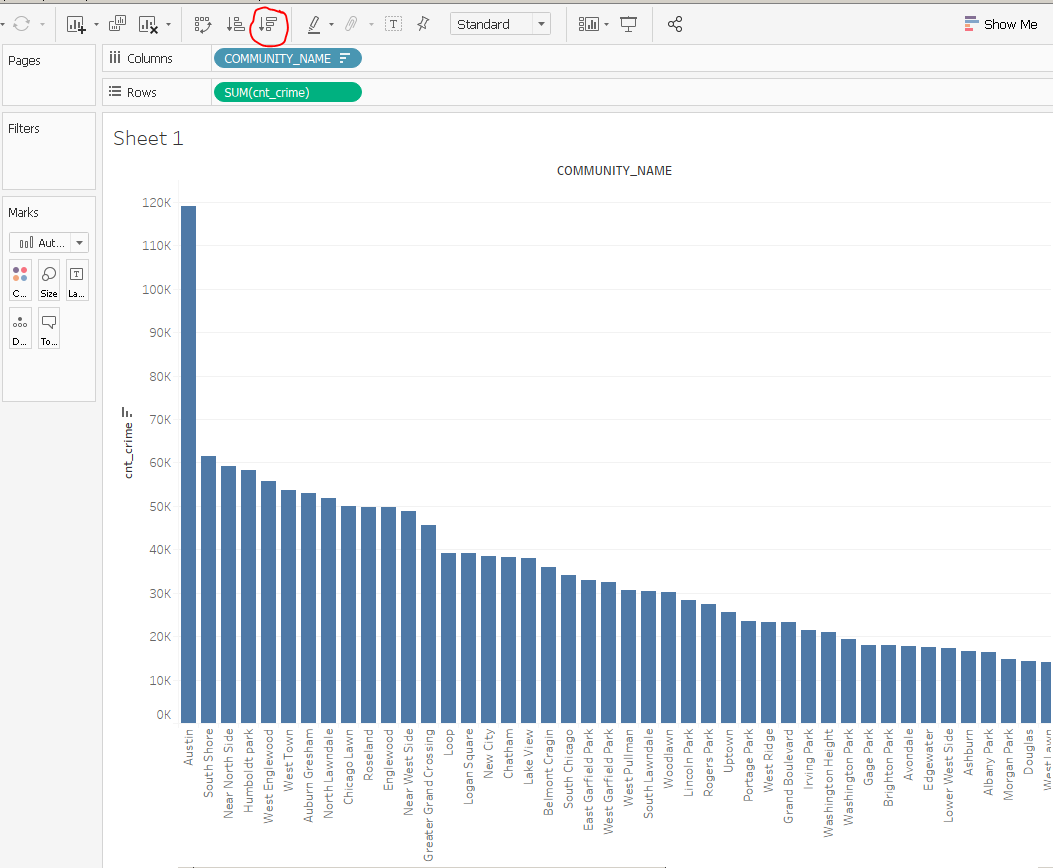


After you done renaming the fields, at the bottom left corner, click “Sheet 1”.

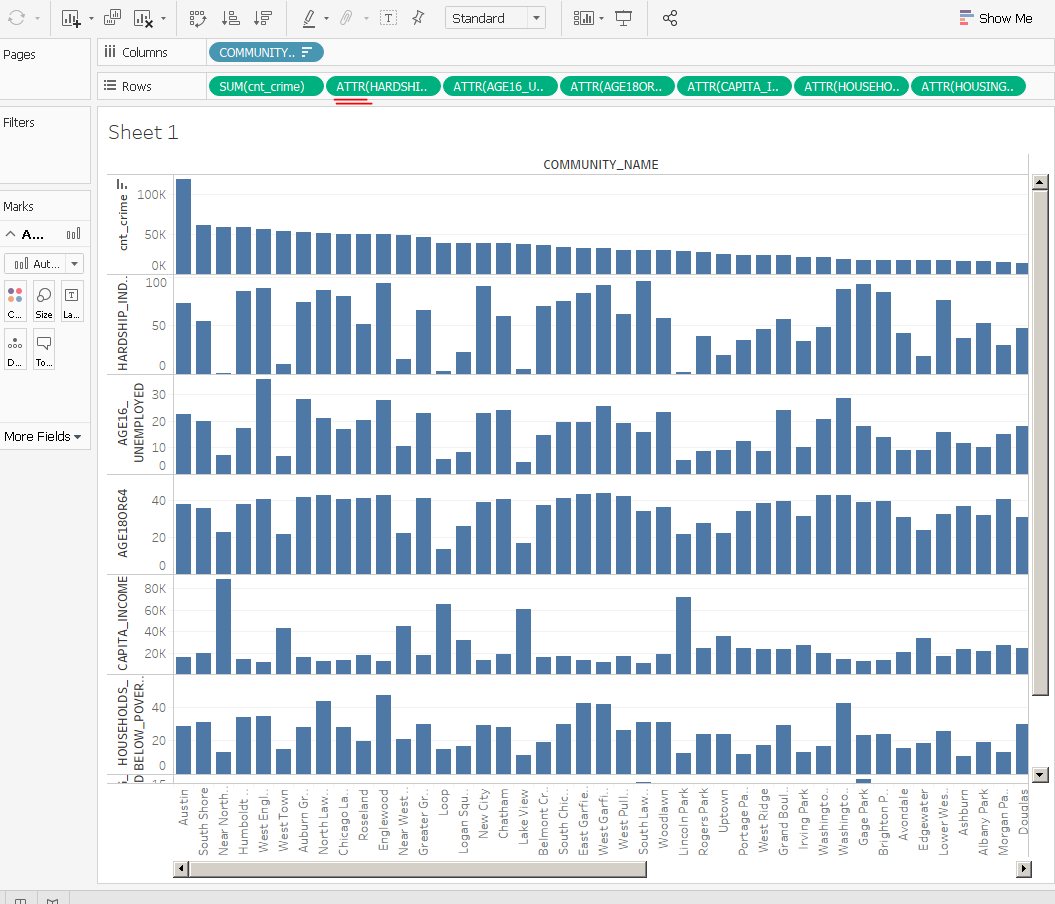
To view how the crime occurrences changes each year from 2008 to 2012 for each community, drag COMMUNITY\_NAME and Year attributes into the Columns box and cnt\_crime as Attribute into Rows box.



To view the highest to lowest crime occurrence from 2008-2012 for each community, drag community\_name into Columns box and SUM of cnt\_crime into Rows box. Above the community\_name field box, click on the button to sort Community\_name descending by cnt\_crime.



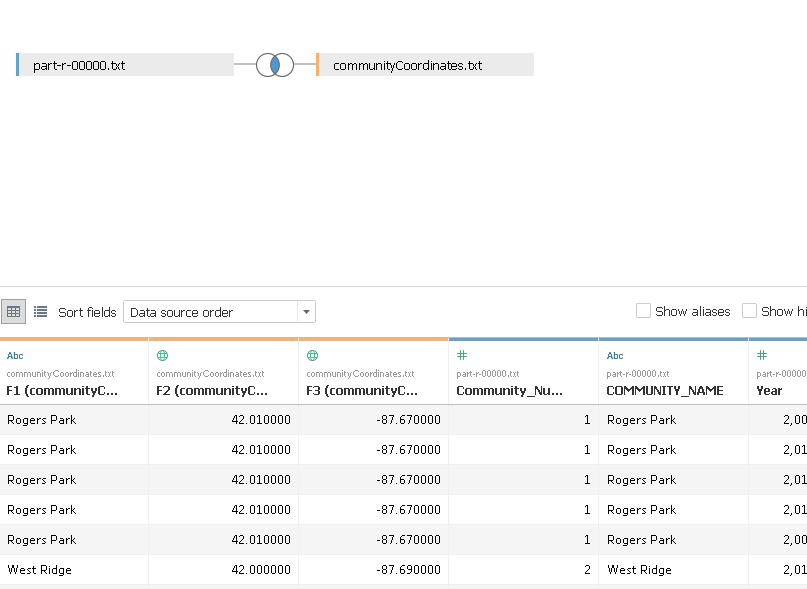
To view how other attributes such as CAPITA\_INCOME, HARDSHIP\_INDEX correlates to the crime occurrences in each community, drag the field as ATTR to the Rows box.



By looking at the charts, there seem to be no correlation between the crime occurrences and any of the social economic attributes.

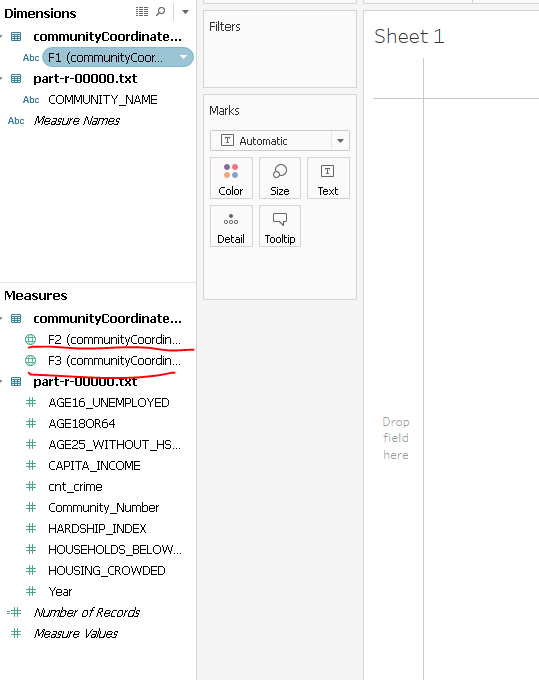
**Geo-Spatial Visualization**

To visualize our data on the map, we pulled the coordinates associated with the community areas from <https://en.wikipedia.org/wiki/Community_areas_in_Chicago> using a Python script and saved them in communityCoordinates.txt and inner joined the file with part-r-00000.txt.

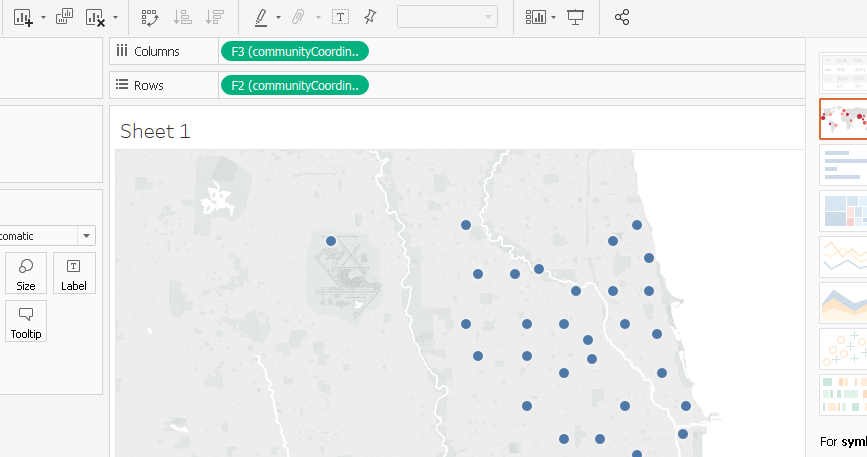


F2 represents the latitude; F3 represents the longitude.

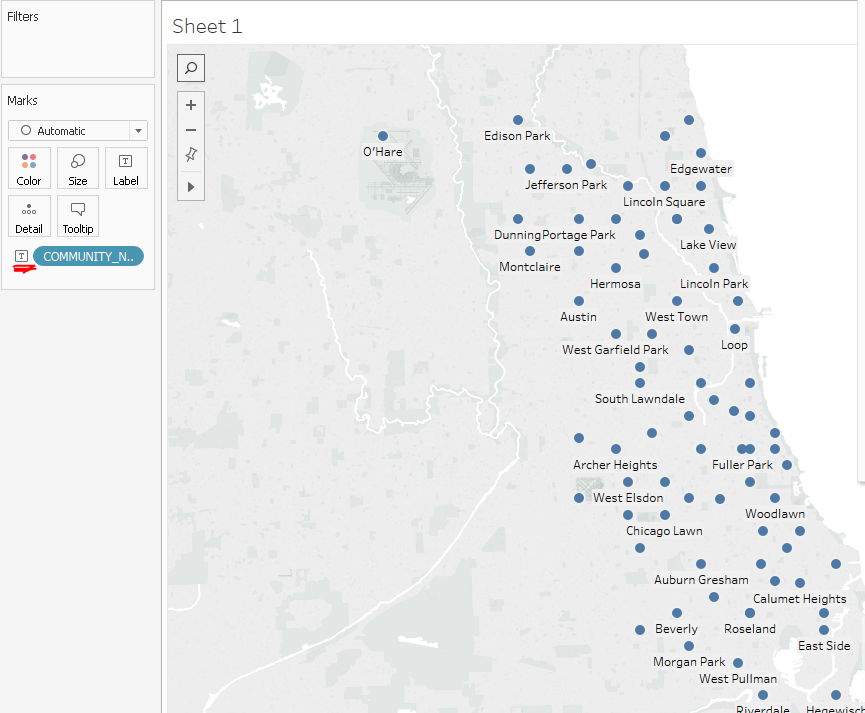
Once both files are joint, go to sheet 1 and locate Measure tab. Change F2’s geographic role to latitude and F3’s to longitude.



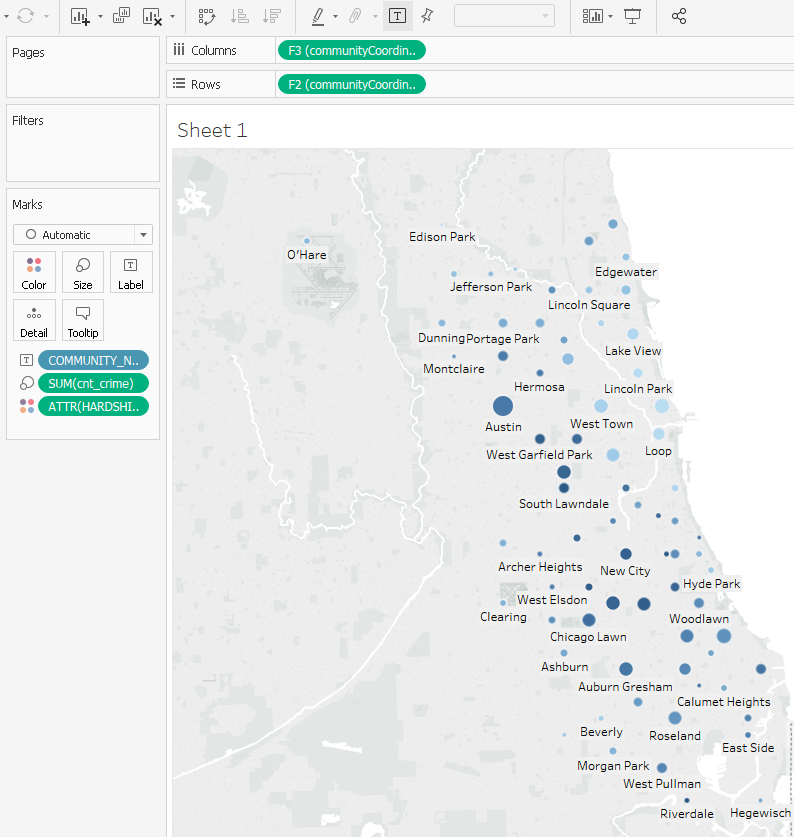
Drag F3 to Columns box and F2 to Rows box. When you drag them in the boxes, their property might change to AVG, make sure to change both F2’s and F3’s property to dimension.



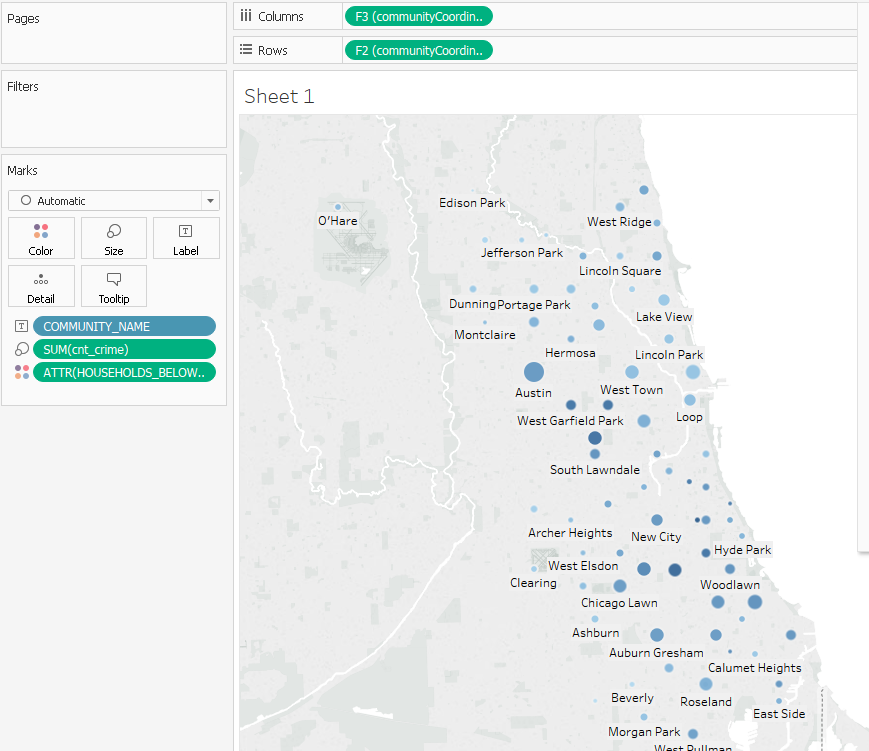
To have the dots on the map labeled, drag COMMUNITY\_NAME to Marks box and label COMMUNITY\_NAME as Detail. This will label the dots as followed:



To compare cnt\_crime with each social economic attribute, drag cnt\_crime and a social economic attribute to the Marks box. Make sure cnt\_crime’s property is SUM and the social economic attribute’s property is an attribute. Label cnt\_crime as size and the social economic attribute as color. The following will be a comparison for cnt\_crime and HARDSHIP\_INDEX.

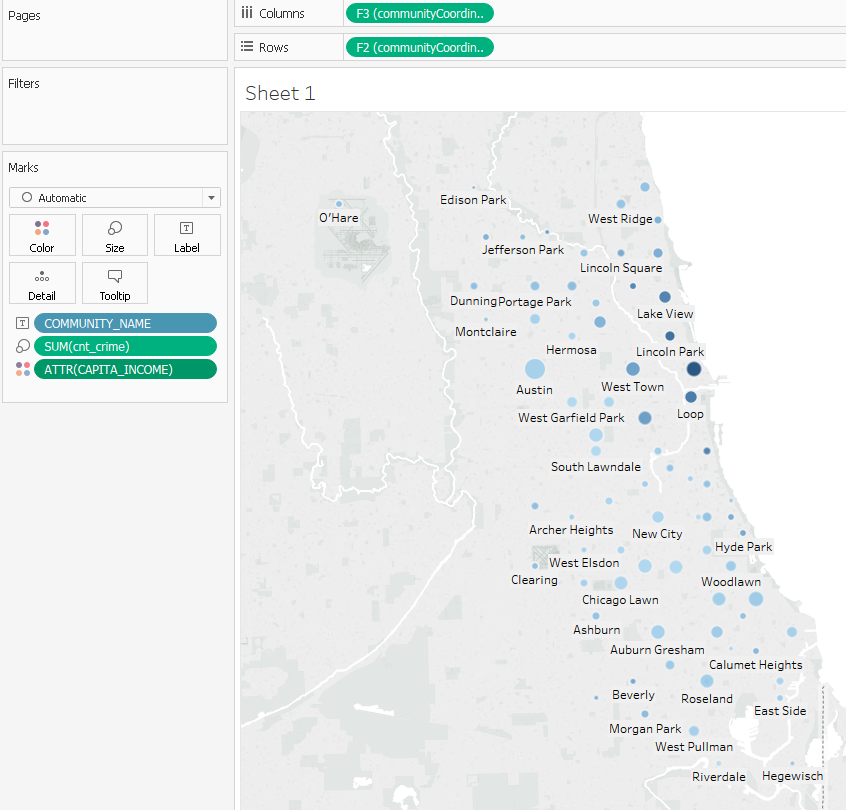
 From looking at the map, we noticed that high crime occurrences usually occurred more toward West side of Chicago. We can see this by the size of the circle of each community. Bigger circle indicates higher crimes. We also noticed that HARDSHIP\_INDEX is typically higher on the West side as well by the intensity of the color. The darker the color is, the higher the index.

Next, we compare cnt\_crime with HOUSEHOLDS\_BELOW\_POVERTY.



The map seems to show that HOUSEHOLDS\_BELOW\_POVERTY is more common on the West side.

Next, we compare cnt\_crime with CAPITA\_INCOME.



By looking at the map, CAPITA\_INCOME is generally higher in the North-East side of Chicago.

Conclusion

By looking at the bar graphs, it was hard to see the correlations between crime rates and the social economic attributes. However, as we moved to geospatial visualization, the link among crime rates and the social economic attributes become visible. Our conclusion is that social economic may play some parts in the crime rates.

References

* 1. <https://catalog.data.gov/dataset/crimes-2001-to-present-398a4>
  2. <https://catalog.data.gov/dataset/census-data-selected-socioeconomic-indicators-in-chicago-2008-2012-36e55>
  3. <https://en.wikipedia.org/wiki/Community_areas_in_Chicago>
  4. <https://github.com/thuh1030/BigDataTeam3>