Chicago’s Crime Data Analysis

using Hadoop, Apache Pig, and Tableau

Raymond Ko  
Diana Hoang  
Jocelyn Gallo  
Luis Cardenas  
Eric Kha

Department of Information Systems, California State University

Los Angeles

Tel. 714-785-7527, Fax. 323-343--5209

e-mail : school@caltstatela.edu

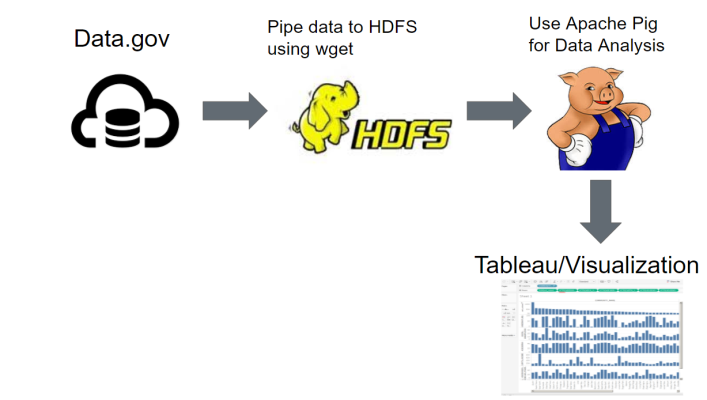
**Abstract:** Our project analyzed the relationship between the crimes that occurred in the City of Chicago and the socioeconomic status of each community area. Does better socioeconomic status mean more or less crime? The data was extracted from the Chicago Police Department’s CLEAR system and U.S. Census Bureau. We compared the two datasets and found a correlation between the socioeconomic status and the crime rate occurred in the City of Chicago. The mediums used to analyze the data included Pig and Tableau. Through the use of Tableau's interface, we noticed an unusual amount of high crime occurrences towards the west side of Chicago. The hardship index, which is a combination of: the percent of crowded housing, the percent of households below poverty, unemployment rate, per capita income, percent aged 25 without a high school diploma, and percent aged under 18 or 65, was the main measure used to finalize our conclusion.

In Chicago, an increase in crime rates correlated with a decrease in socioeconomic status.

**1. Introduction**

The city of Chicago has some of the strictest gun control laws. In fact, legally obtaining guns is a tedious process and requires the signing of numerous documents for protective purposes. We wanted to take a different approach and instead of correlating guns with crime rates, we wanted to dig a little deeper. Fortunately, the datasets were already provided by Chicago’s police department and the U.S. Census Bureau. Upon extensive research, we weren’t as surprised to see the occurrences of crimes in the cities of Chicago placed amongst the top five. As we tweaked certain attributes and variables to find a correlation, a relation was found upon studying a dataset with the communities of Chicago. Through the following research, our dataset illustrates that an increase in crime rates portrays an inverse relationship with areas with lower socioeconomic status.

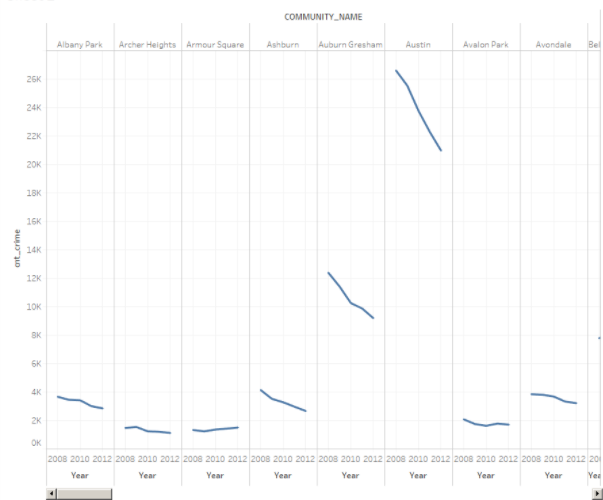
**2. General Tools**

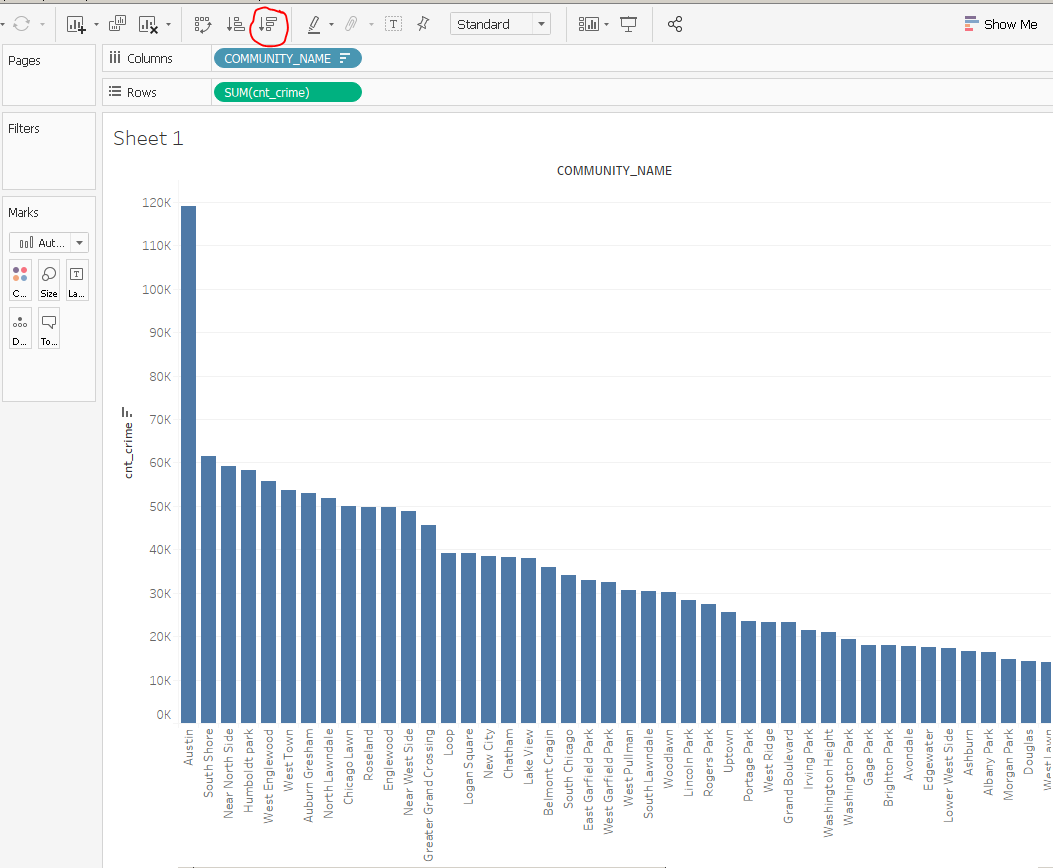
****

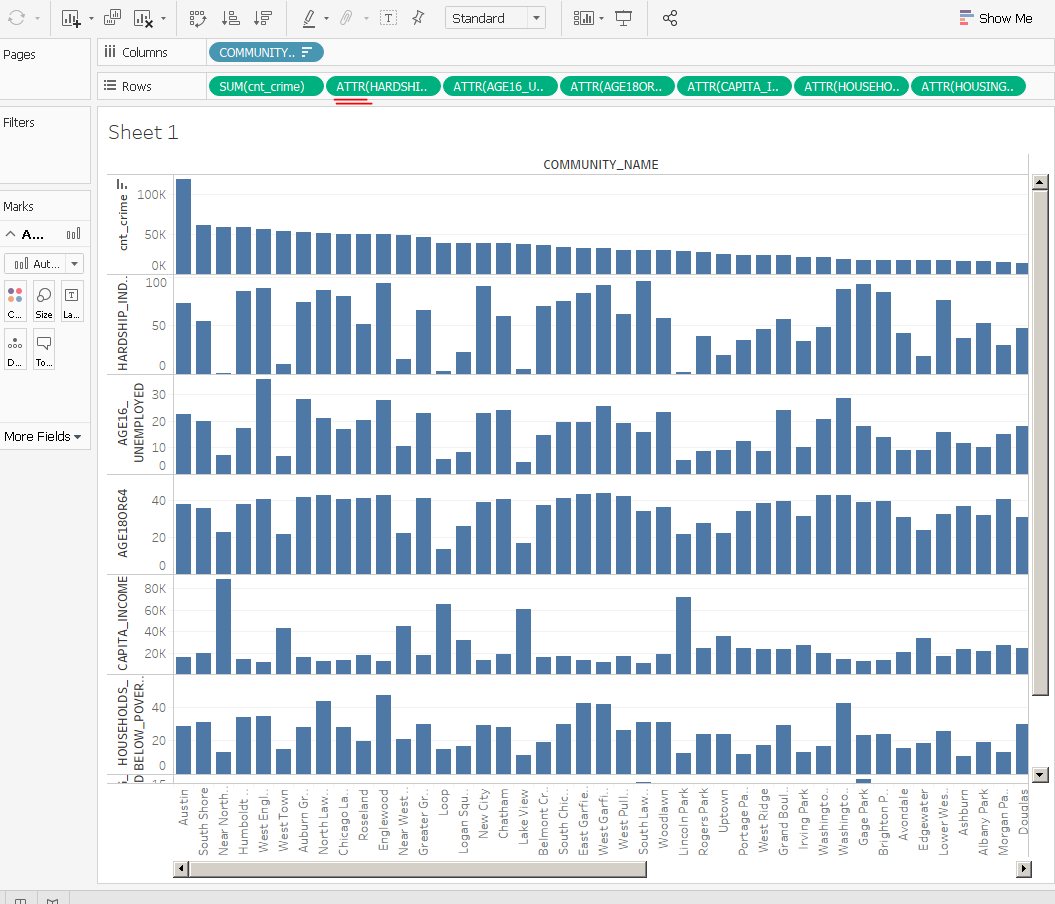
The following research used HDFS, Pig, and tableau to output data. The mode of analyzing the data was through IBM bluemix trial account. Through the use of the account, members obtain access to two nodes and unlimited data processing for over one hundred fifty days. The data files crime.csv and social-econ.csv were first downloaded using wget command in HDFS, and then uploaded to the HDFS file system. Data was later loaded into Pig for analysis. Pig has various means to analyze huge clusters of data without overwhelming a system. The working data sums up to 1.4 gigabytes and analyzing it traditionally may take a cumbersome amount of time. Fortunately, Pig can organize the data rather quickly. Through the use of various scripts in Pig, the data is normalized to filter needed information. To visualize the crime rates, Tableau will craft the data into histograms and map charts. Crime occurrences and the respective correlations were found through these mediums. Out platform specs contained one management mode using a 4 core VCPU with 24 gigabytes in our data node. The files were uploaded onto hdfs and using this set up.

**2.1 Crime Occurrences in all Cities**

All of the data was extracted from the **U.S. C**ensus Bureau and Chicago Police Department. The data shows that Austin has the highest amount of crime. Cities in Chicago do follow such as South Shore and Near North in second and third place respectively. Figure 1.1 shown below. We used Hadoop and tableau to output and analyze the above data. Hadoop has various means to analyze big data without overwhelming a system. Tableau visualizes the data after the data has been normalized and edited. The findings above illustrate a general overview of the crime occurrences. To add substance to our data, we added the hardship index, which is a combination of; the percent of crowded housing, the percent of households below poverty, unemployment rate, per capita income, percent aged 25 without a high school diploma, and percent aged under 18 or 65.

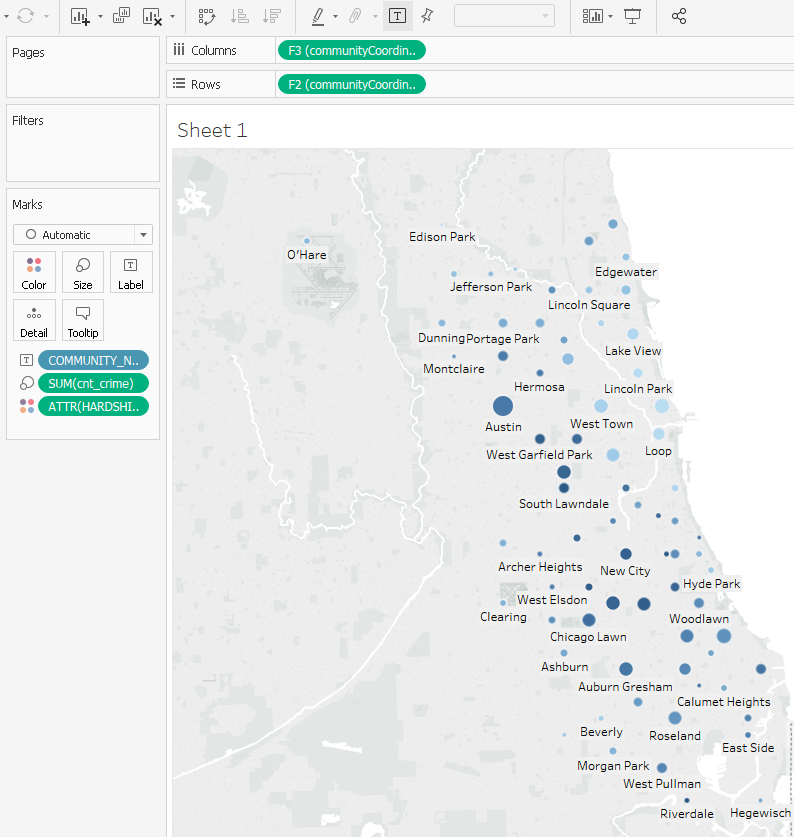
  
Figure 1.1 The changes in crime rate for each community from 2008-2012.

  
Figure 1.2 Crime occurrences in all communities.

Figure 1.3 Comparisons of crime occurrences with other socio-economic attributes such as hardship index, crowded housing, the percent of households below poverty, and etc. for each community.

In Figure 1.1, we can see that the crime rate has been decreasing for each community from year 2008 to 2012. Upon adding the hardship index and other attributes, meaningful data is yet to be discovered. A strong correlation between the socio economic attributes and the crime occurrences of each city is yet to be found. However, we were able to find the crime occurrences from each year using 2008 through 2012 data. To do so, we used various attributes such as the community variable name, year attributes, and count of crime into separate columns. In other words, the column name consisted of the community\_name and the year while the rows tabulated the count of crime. The highest count for crime occurred during 2008 in Austin whereas Chicago trails not too far. However, we still need to dig a little deeper to draw further conclusions.

**3. Findings**

A couple of more manipulations were needed in order to derive a final conclusion. Upon pulling coordinates associated with the community areas and creating an association with our research files, a new discovery was made. The two attributes cnt\_crime and hardship\_index were the variables used to test any findings. In figure three, higher crime occurrences usually occur more toward West side of Chicago. We can see this by the size of the circle of each community. Bigger circle indicate higher crime occurrences. 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. As mentioned earlier, the hardship\_index combines several factors to indicate the poverty levels of a particular area. The manipulation of the data attributes paint a clear trend in regards to crime occurrences. Low socioeconomic status does have a relation to higher crime occurrences in Chicago.  
  
Figure 3 Crime occurrences in relation with the hardship index in Chicago for each community.

**4. Conclusion**

We have concluded that a possible correlation exists between socioeconomic level and crime occurrences. The methods and standards we used to manipulate the data weren’t too difficult. However, coming up with a medium to measure the data we exported would be the next step in drawing further conclusions. With this said, it appears that Chicago is still a city with high crime counts. Instead of manipulating stricter gun laws and the mode of violence, perhaps a solution to fix crime rate lies in manipulating poverty levels. The variables of poverty are beyond the scope of this research but it is clear that neighborhoods with lower economic status tend to have higher 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