Big Data Analytics Symposium - Fall 2018

Analytics Project: Quantitative Analysis of Crime Incidents in the City of Chicago, IL using Data Analytics Techniques

Team: Daniel Rivera & Alisha Sawant

Abstract: By combining different data sources, the analytic will extract information that is correlated to the pattern of crime incidents in Chicago. Demographics, train ridership and traffic data, and business licenses in the areas were considered.

Motivation

Who are the users of this analytic? - Police Department of Chicago

Who will benefit from this analytic? - Police Department of Chicago & the residents of Chicago

Why is this analytic important?

This analytic can be used to anticipate the occurrence of certain types of crimes within the community areas of a Chicago. With this information, the police can increase the number of policemen on patrol and the patrol frequency in those areas.

Quantitative Analysis of Crime Incidents in the City of Chicago, IL using Data Analytics Techniques

Goodness

What steps were taken to assess the 'goodness' of the analytic?

Table 1: Pearson correlation between demographic features and crime rate (* indicates significant correlations with p-value less than 5%).

Feature	Correlation	p-value
Total Population	-0.1269	0.2716
Population Density	-0.1972	0.0855
Poverty Index	**0.5573***	1.403e-07
Disadvantage Index	**0.5959***	1.082e-08
Residential Stability	-0.0453	0.6965
Ethnic Diversity	**-0.5545***	1.678e-07
Percentage of Black	**0.6696***	2.779e-11
Percentage of Hispanic	**-0.3820***	0.0006

(Source: Graif C, Kifer D, Li Z, Wang H. Crime Rate Inference with Big Data. Available at https://www.kdd.org/kdd2016/papers/files/adp1044-wangA.pdf)

According to a previous analysis performed on the Chicago Crime dataset, poverty index is one of the demographic factors that exhibits a significant correlation to the number of crime incidents in the city.

Our results exhibit similar characteristics, so we have reason to believe that the analysis is trustworthy.

Data Sources

Name: Crimes - 2001 to present

Description: Reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present.

Size of data: 1.7 GB

Name: Census Community Data

Description: A combination of multiple datasets giving information about age, ethnicity and economic demographics.

Size of data:

Name: Train Data – 'L' Station Entries

Description: Shows daily totals of ridership, by station entry, for each 'L' station dating back to 2001.

Size of data:

Data Sources (contd…)

Name: Chicago Traffic Tracker

Description: Contains the historical estimated congestion for 1270 traffic segments, in selected time periods from August 2011 to May 2018.

Size of data: 640 MB

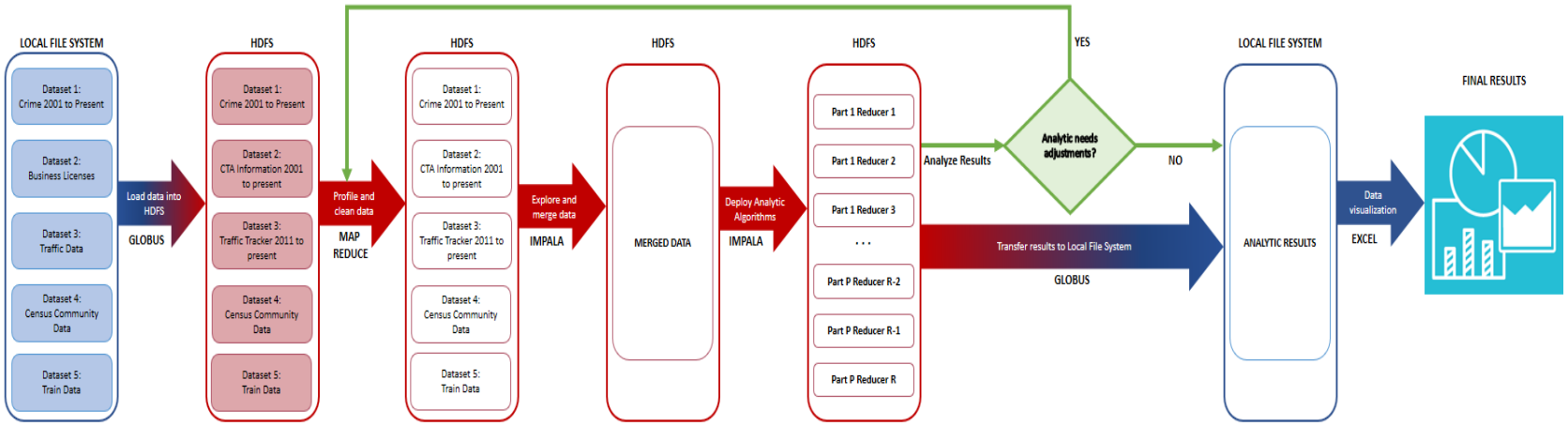
Name: Business Licenses

Description: Business licenses issued by the Department of Business Affairs and Consumer Protection in the City of Chicago from 2002 to the present.

Size of data: 300 MB

Quantitative Analysis of Crime Incidents in the City of Chicago, IL using Data Analytics Techniques

Design Diagram



Platform(s) on which the analytic ran:
NYU HPC Cluster - Dumbo

Results

1. Demographic factors, such as income below poverty level, repeatedly appear as highly correlated to several crime types in most community areas.

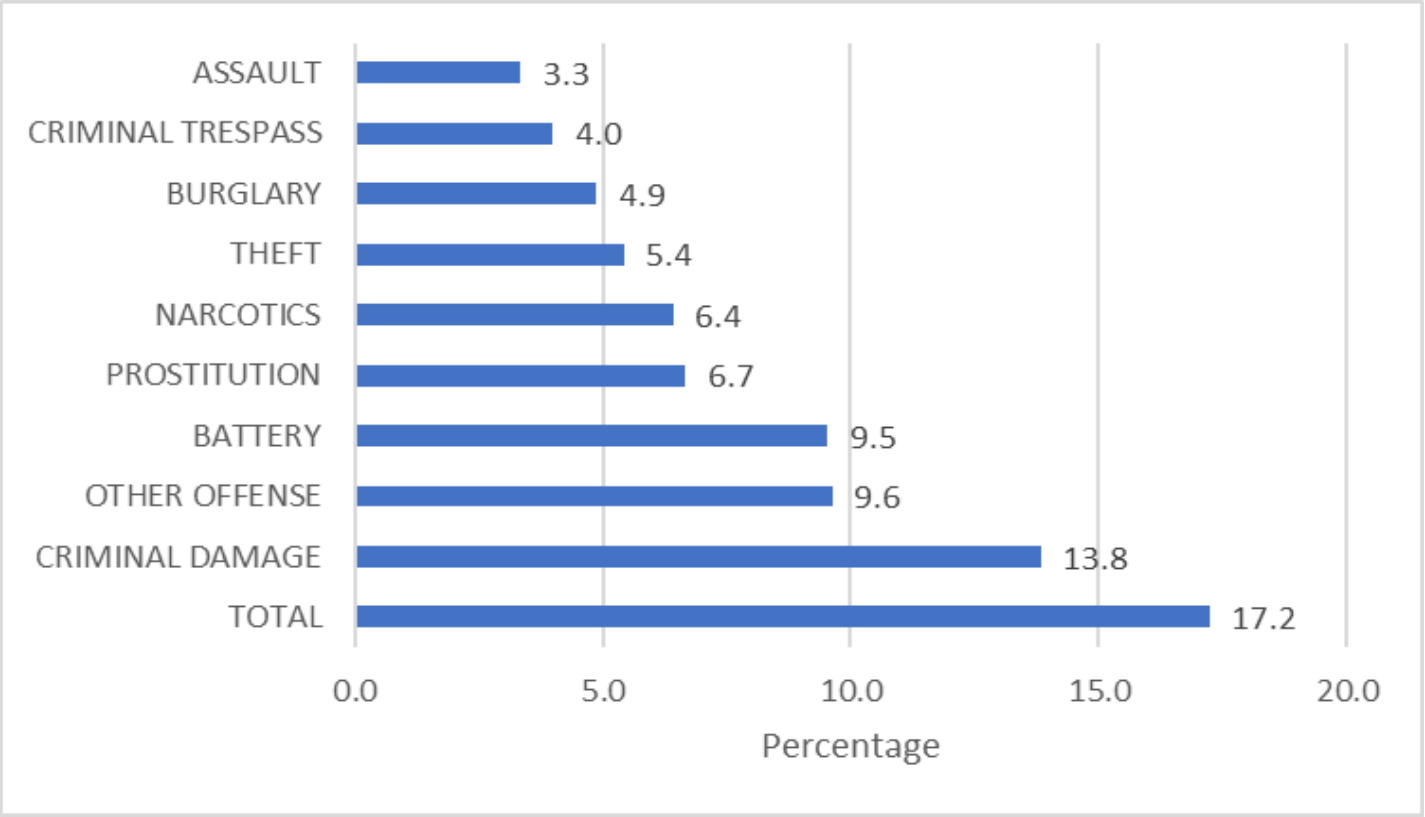
2. The number of public transportation buses within the city exhibits a high correlation with **theft** crime incidents.

3. Certain combinations of demographic factors and community areas exhibit a significant **negative correlation** with crime incidents.

Results

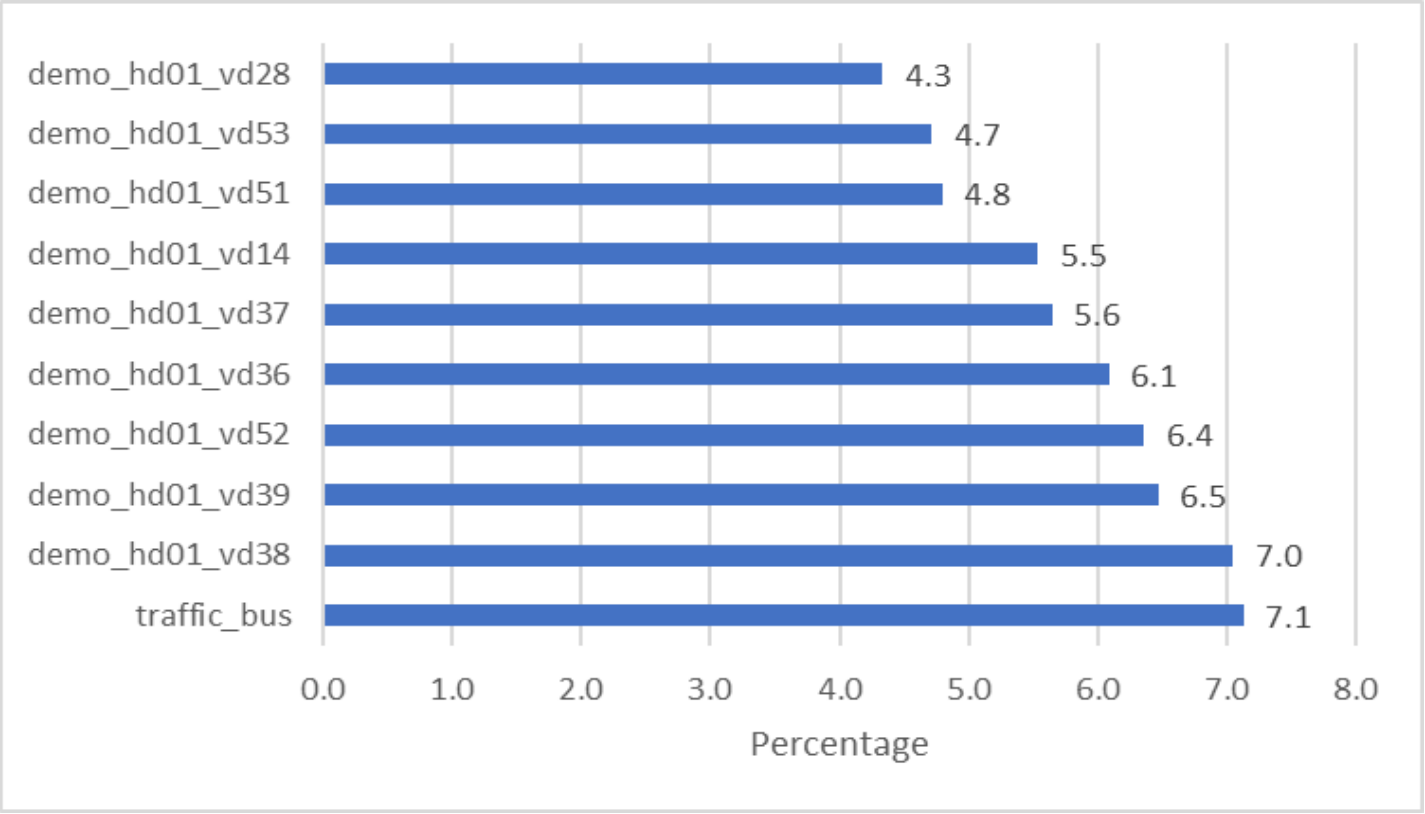
Factor	*Crime Type*	*Community Area*	*Correlation*
Income in the past 12 months below poverty level	*Total*	*00*	*0.8319*
Public transportation buses	*Theft*	*00*	*0.8454*
Average traffic speed	*Prostitution*	*57*	*0.8117*
Income below poverty level: male 55 to 64 years	*Total*	*66*	*-0.8706*
Income below poverty level: female 55 to 64 years	*Total*	*66*	*-0.8600*

Results



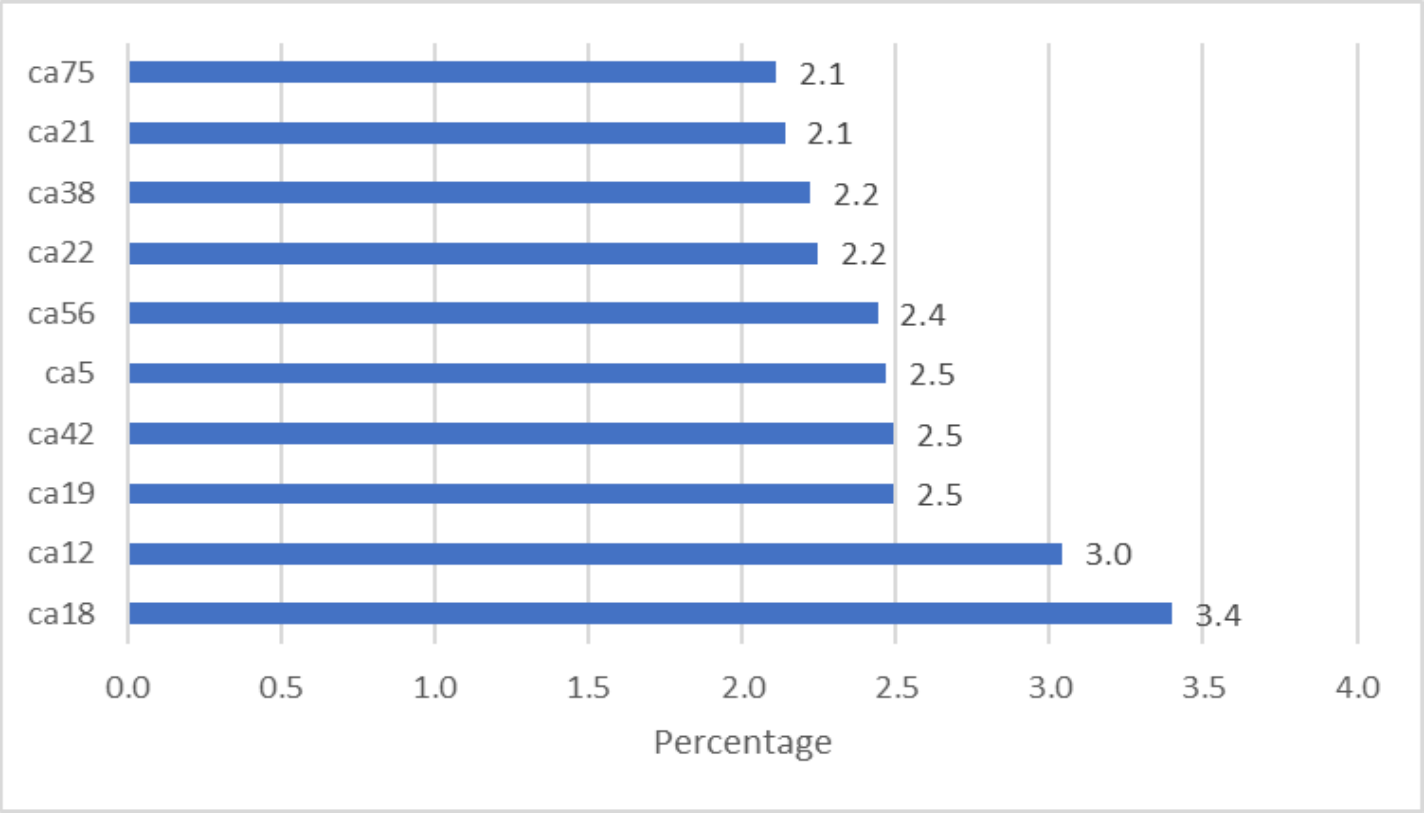
Top 10 Crime Types by Percentage with High Correlations

Results



Top 10 Factors by Percentage with High Correlations

Results



Top 10 Community Areas by Percentage with High Correlations

Obstacles

1. Mapping the location information (latitude, longitude) to the corresponding community areas.
   - Using the Socrata Open Data API (SODA) we were able to query a dataset that includes the geometry of all the community areas:

```
String w = String.format("intersects(the_geom, '%s')", point);
SoqlQuery query = new SoqlQueryBuilder()
                      .addSelectPhrase("area_numbe")
                      .setWhereClause(w)
                      .build();
```

   - Connecting to the API was a time-consuming process, so making use of the parallelized computation capabilities of MapReduce was crucial to complete the task within a reasonable time (next slide)

Quantitative Analysis of Crime Incidents in the City of Chicago, IL using Data Analytics Techniques

Obstacles

Job Overview

Job Name:	Soda Job Split
State:	RUNNING
Uberized:	false
Started:	Wed Nov 14 01:11:05 EST 2018
Elapsed:	18mins, 26sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Wed Nov 14 01:11:03 EST 2018	compute-1-4.local:8042	logs

Task Type	Progress	Total	Pending	Running	Complete
Map		302	0	0	302
Reduce		0	0	0	0

Attempt Type	New	Running	Failed	Killed	Successful
Maps	0	0	0	0	302
Reduces	0	0	0	0	0

Summary

- Crime is influenced by multiple factors that may be complex and unsuspected.
- Positively correlated factors: take action towards minimizing them.
- Negatively correlated factors: take action towards extrapolating them to other areas.
- Future work: time series for causality analysis and forecasting.

References

1. Wang, Hongjian, et al. "Crime rate inference with big data." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
2. Zhao, Xiangyu, and Jiliang Tang. "Crime in Urban Areas:: A Data Mining Perspective." ACM SIGKDD Explorations Newsletter 20.1 (2018): 1-12.
3. Williams, Matthew L., Pete Burnap, and Luke Sloan. "Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns." The British Journal of Criminology 57.2 (2017): 320-340.
4. Pramanik, Md Ileas, et al. "A framework for criminal network analysis using big data." e-Business Engineering (ICEBE), 2016 IEEE 13th International Conference on. IEEE, 2016.

Thank you!