Big Data Analytics Symposium - Fall 2022

Analytics Project: Chicago Crime & Community Analysis

Team Members:

- Astha Gupta (ag7982)
- Arpan Ghoshal (ag8821)
- Omar Benkraouda (omb244)

Abstract

When using multiple data sources, we are able to come to different compelling insights, correlations, and conclusions regarding Crime in Chicago. Traffic and train data, Demographics (age, income, etc..), as well as crime types were utilized to draw conclusions. Hive was used to process queries and store data and Tableau was used to plot and visualize our correlation graphs.

Platform(s) where the application runs: NYU Dataproc Cluster.

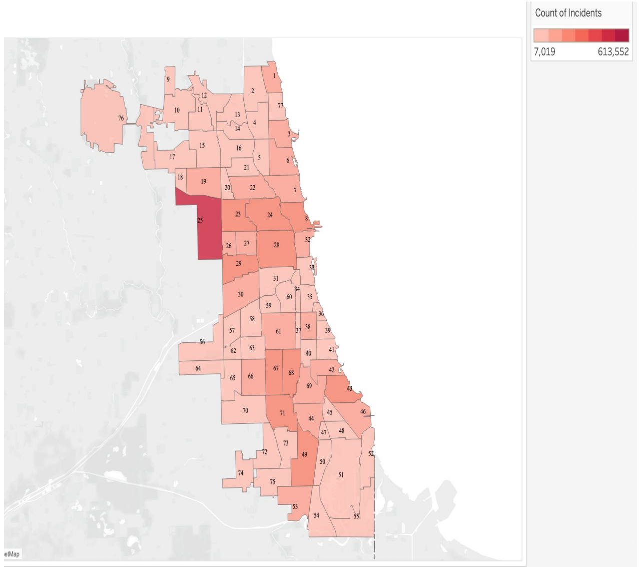


Figure 1: Number of crimes in community areas

Motivation

Who are the users of this analytic?
Police Department of Chicago, City of Chicago

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

Why is this analytic important?

This analytic can be used to anticipate crime spikes and increase patrolling and governance in areas with higher crime rates. It can also help in informing decisions regarding policies for each area.

Goodness

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.

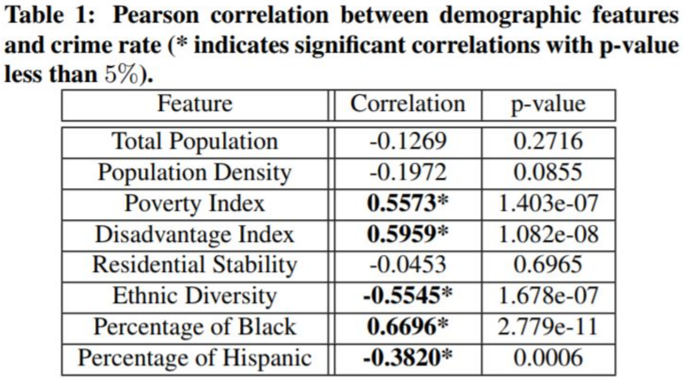


Table 1: Pearson correlation between demographic features and crime rate (* indicates significant correlations with p-value less than 5%).

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

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: 4 KB

Name: Train Data – ‘L’ Station Entries

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

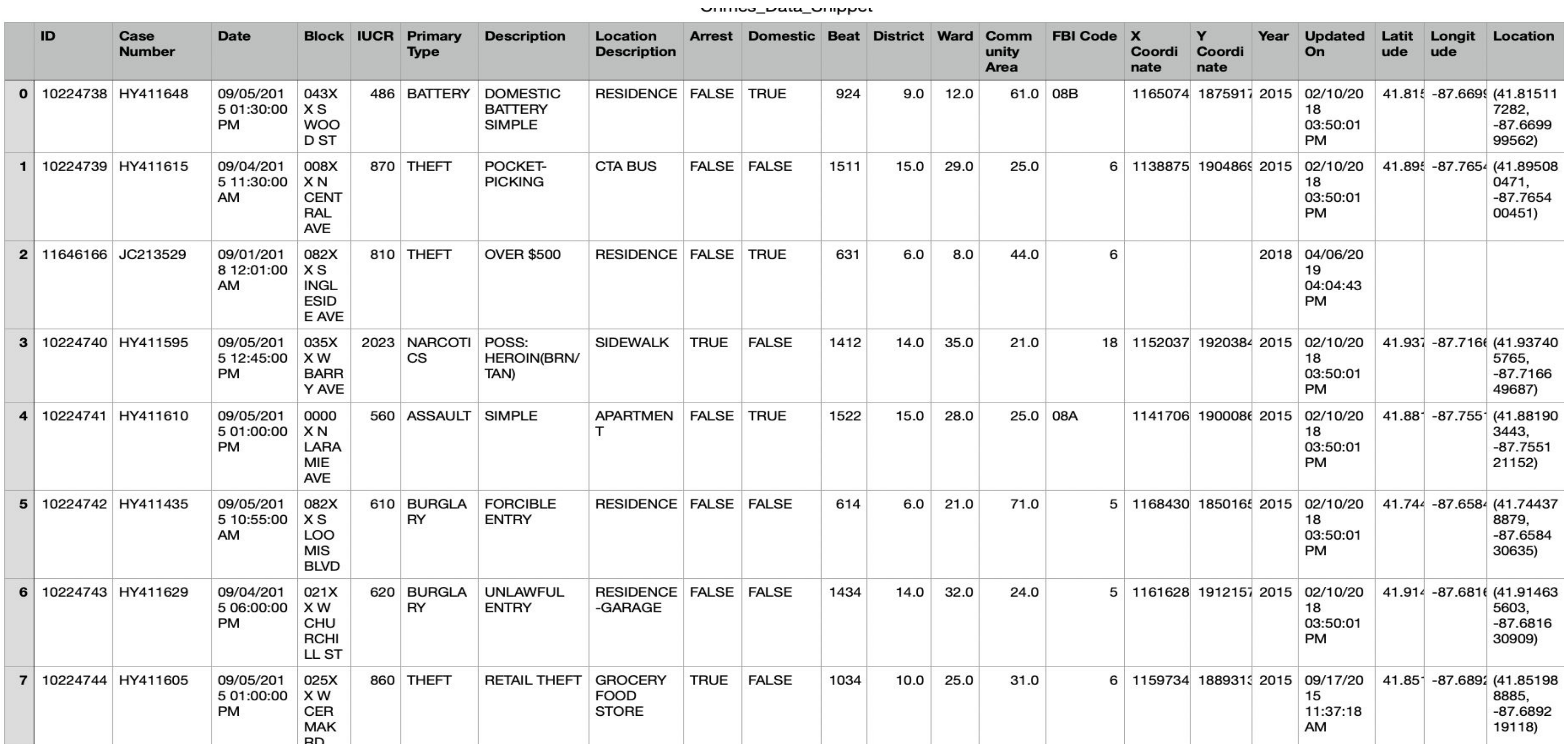
Size of data: 41MB

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

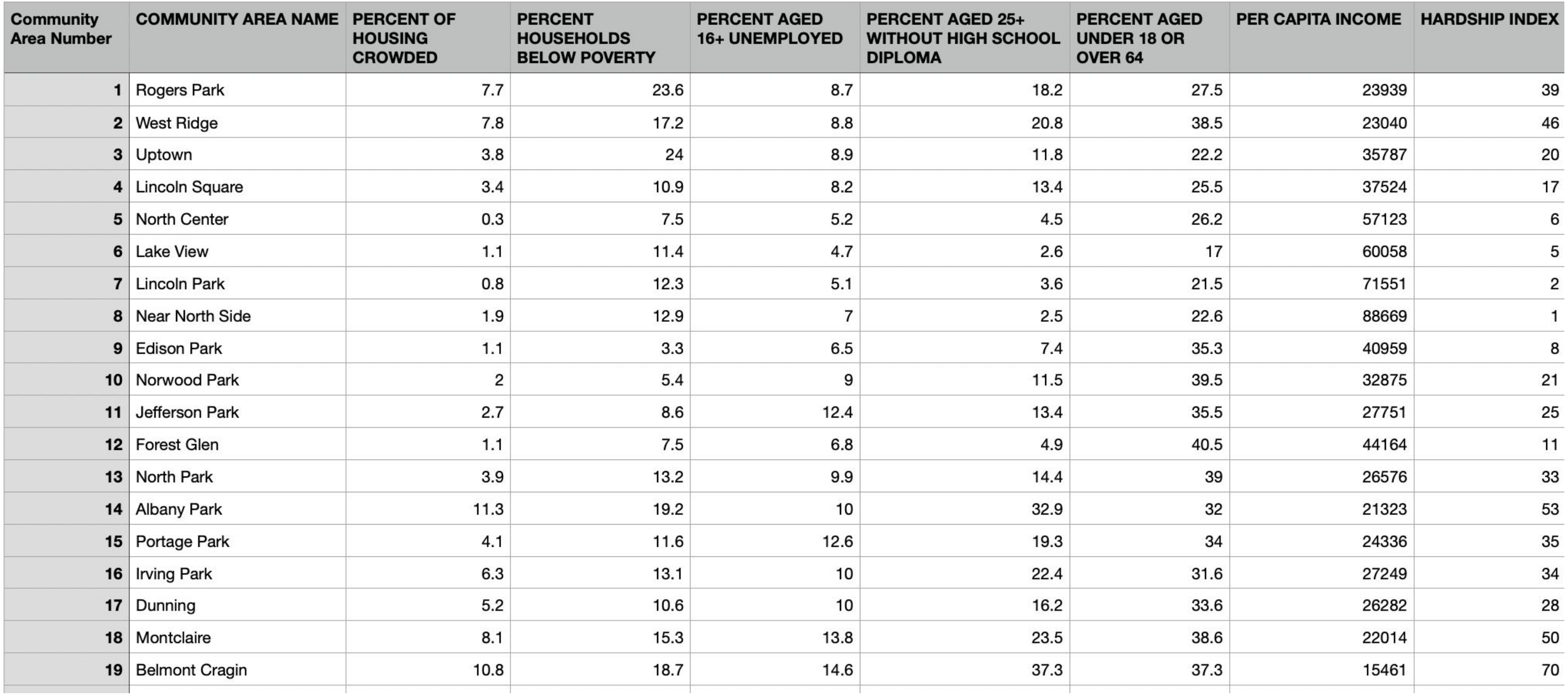
Data Sample: Chicago Crimes Data



Crimes_Data_Snippet

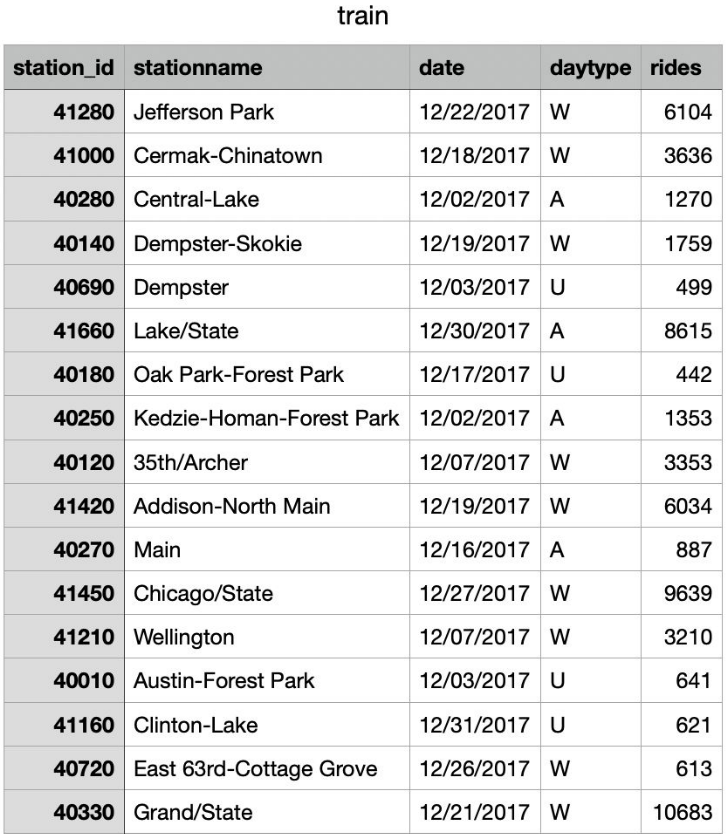
	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
0	10224738	HY411648	09/05/2015 01:30:00 PM	043XX S WOOD ST	486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	FALSE	TRUE	924	9.0	12.0	61.0	08B	1165074	1875917	2015	02/10/2018 03:50:01 PM	41.815	-87.6699	(41.815117282, -87.669999562)
1	10224739	HY411615	09/04/2015 11:30:00 AM	008XX N CENTRAL AVE	870	THEFT	POCKET-PICKING	CTA BUS	FALSE	FALSE	1511	15.0	29.0	25.0	6	1138875	1904869	2015	02/10/2018 03:50:01 PM	41.895	-87.7654	(41.895080471, -87.765400451)
2	11646166	JC213529	09/01/2018 12:01:00 AM	082XX S INGLESIDE AVE	810	THEFT	OVER $500	RESIDENCE	FALSE	TRUE	631	6.0	8.0	44.0	6			2018	04/06/2019 04:04:43 PM			
3	10224740	HY411595	09/05/2015 12:45:00 PM	035XX W BARRY AVE	2023	NARCOTICS	POSS: HEROIN(BRN/TAN)	SIDEWALK	TRUE	FALSE	1412	14.0	35.0	21.0	18	1152037	1920384	2015	02/10/2018 03:50:01 PM	41.937	-87.7166	(41.937405765, -87.716649687)
4	10224741	HY411610	09/05/2015 01:00:00 PM	0000X N LARAMIE AVE	560	ASSAULT	SIMPLE	APARTMENT	FALSE	TRUE	1522	15.0	28.0	25.0	08A	1141706	1900086	2015	02/10/2018 03:50:01 PM	41.881	-87.7551	(41.881903443, -87.755121152)
5	10224742	HY411435	09/05/2015 10:55:00 AM	082XX S LOOMIS BLVD	610	BURGLARY	FORCIBLE ENTRY	RESIDENCE	FALSE	FALSE	614	6.0	21.0	71.0	5	1168430	1850165	2015	02/10/2018 03:50:01 PM	41.744	-87.6584	(41.744378879, -87.658430635)
6	10224743	HY411629	09/04/2015 06:00:00 PM	021XX W CHURCHILL ST	620	BURGLARY	UNLAWFUL ENTRY	RESIDENCE-GARAGE	FALSE	FALSE	1434	14.0	32.0	24.0	5	1161628	1912157	2015	02/10/2018 03:50:01 PM	41.914	-87.6816	(41.914635603, -87.681630909)
7	10224744	HY411605	09/05/2015 01:00:00 PM	025XX W CERMAK RD	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	TRUE	FALSE	1034	10.0	25.0	31.0	6	1159734	1889313	2015	09/17/2015 11:37:18 AM	41.851	-87.6892	(41.851988885, -87.689219118)

Data Sample: Socioeconomic Indicators



Community Area Number	COMMUNITY AREA NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX
1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39
2	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46
3	Uptown	3.8	24	8.9	11.8	22.2	35787	20
4	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17
5	North Center	0.3	7.5	5.2	4.5	26.2	57123	6
6	Lake View	1.1	11.4	4.7	2.6	17	60058	5
7	Lincoln Park	0.8	12.3	5.1	3.6	21.5	71551	2
8	Near North Side	1.9	12.9	7	2.5	22.6	88669	1
9	Edison Park	1.1	3.3	6.5	7.4	35.3	40959	8
10	Norwood Park	2	5.4	9	11.5	39.5	32875	21
11	Jefferson Park	2.7	8.6	12.4	13.4	35.5	27751	25
12	Forest Glen	1.1	7.5	6.8	4.9	40.5	44164	11
13	North Park	3.9	13.2	9.9	14.4	39	26576	33
14	Albany Park	11.3	19.2	10	32.9	32	21323	53
15	Portage Park	4.1	11.6	12.6	19.3	34	24336	35
16	Irving Park	6.3	13.1	10	22.4	31.6	27249	34
17	Dunning	5.2	10.6	10	16.2	33.6	26282	28
18	Montclaire	8.1	15.3	13.8	23.5	38.6	22014	50
19	Belmont Cragin	10.8	18.7	14.6	37.3	37.3	15461	70

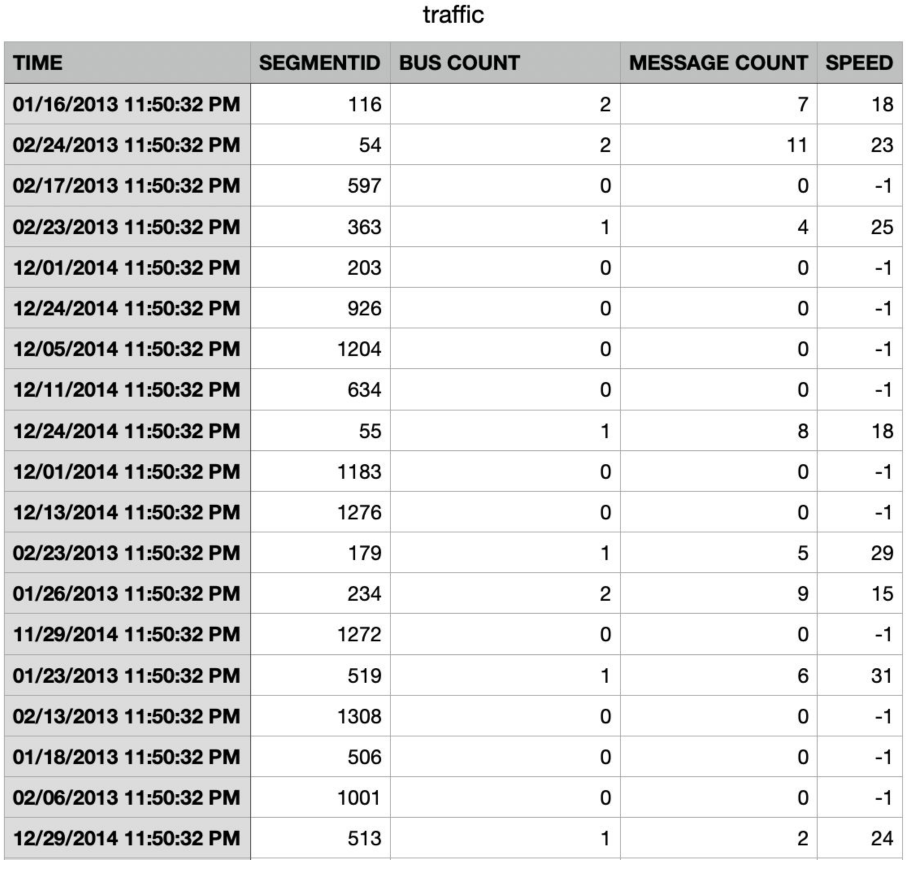
Data Sample - Train Data – ‘L’ Station Entries



train

station_id	stationname	date	daytype	rides
41280	Jefferson Park	12/22/2017	W	6104
41000	Cermak-Chinatown	12/18/2017	W	3636
40280	Central-Lake	12/02/2017	A	1270
40140	Dempster-Skokie	12/19/2017	W	1759
40690	Dempster	12/03/2017	U	499
41660	Lake/State	12/30/2017	A	8615
40180	Oak Park-Forest Park	12/17/2017	U	442
40250	Kedzie-Homan-Forest Park	12/02/2017	A	1353
40120	35th/Archer	12/07/2017	W	3353
41420	Addison-North Main	12/19/2017	W	6034
40270	Main	12/16/2017	A	887
41450	Chicago/State	12/27/2017	W	9639
41210	Wellington	12/07/2017	W	3210
40010	Austin-Forest Park	12/03/2017	U	641
41160	Clinton-Lake	12/31/2017	U	621
40720	East 63rd-Cottage Grove	12/26/2017	W	613
40330	Grand/State	12/21/2017	W	10683

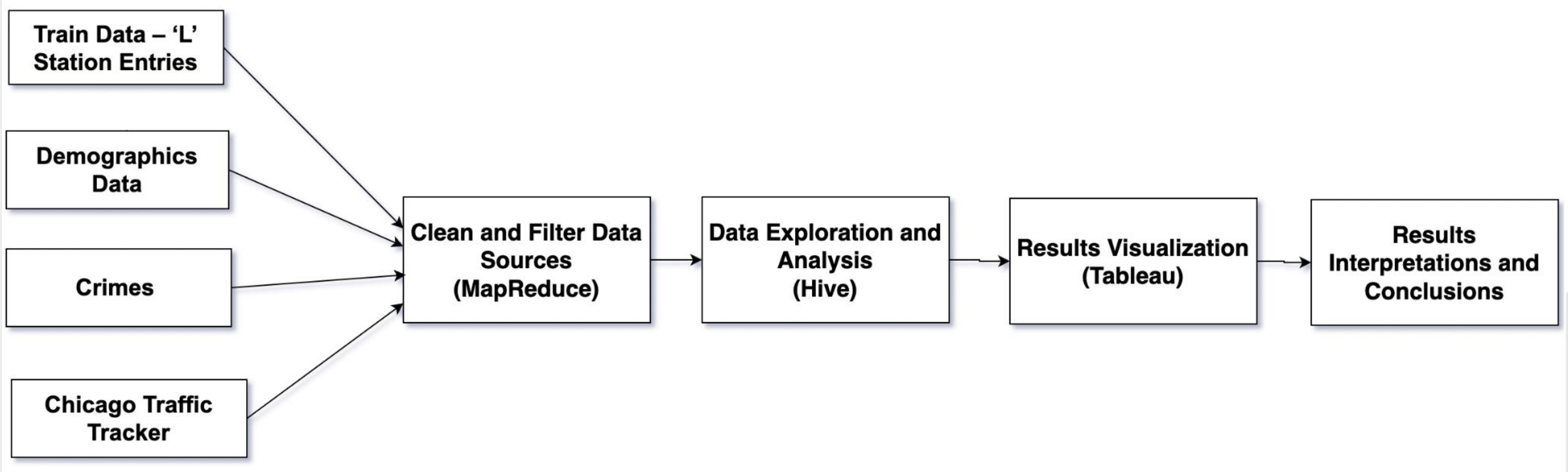
Data Sample - Chicago Traffic Tracker



traffic

TIME	SEGMENTID	BUS COUNT	MESSAGE COUNT	SPEED
01/16/2013 11:50:32 PM	116	2	7	18
02/24/2013 11:50:32 PM	54	2	11	23
02/17/2013 11:50:32 PM	597	0	0	-1
02/23/2013 11:50:32 PM	363	1	4	25
12/01/2014 11:50:32 PM	203	0	0	-1
12/24/2014 11:50:32 PM	926	0	0	-1
12/05/2014 11:50:32 PM	1204	0	0	-1
12/11/2014 11:50:32 PM	634	0	0	-1
12/24/2014 11:50:32 PM	55	1	8	18
12/01/2014 11:50:32 PM	1183	0	0	-1
12/13/2014 11:50:32 PM	1276	0	0	-1
02/23/2013 11:50:32 PM	179	1	5	29
01/26/2013 11:50:32 PM	234	2	9	15
11/29/2014 11:50:32 PM	1272	0	0	-1
01/23/2013 11:50:32 PM	519	1	6	31
02/13/2013 11:50:32 PM	1308	0	0	-1
01/18/2013 11:50:32 PM	506	0	0	-1
02/06/2013 11:50:32 PM	1001	0	0	-1
12/29/2014 11:50:32 PM	513	1	2	24

Design Diagram


Train Data – ‘L’ Station Entries
Demographics Data
Crimes
Chicago Traffic Tracker
Clean and Filter Data Sources (MapReduce)
Data Exploration and Analysis (Hive)
Results Visualization (Tableau)
Results Interpretations and Conclusions

Challenge

Different datasets had different time periods. Other data not available.

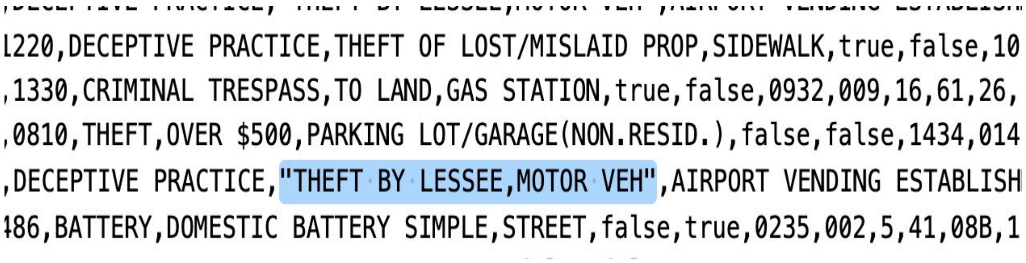
- Crimes data was available from 2001 to present.
- Socioeconomic Factors data was for a period from 2008 to 2012.
- Traffic Data was available from 2011 to 2018.

Thus, we had to find ways to work with the datasets we had.

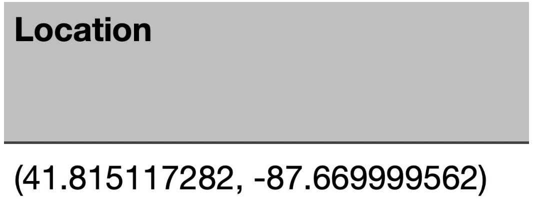
We independently analysed each data with the Crimes data for that period.

Challenge

Datasets had comma separated values within fields, and double quotes which caused problems in reading data.



```
1220,DECEPTIVE PRACTICE,THEFT OF LOST/MISLAID PROP,SIDEWALK,true,false,10
,1330,CRIMINAL TRESPASS,TO LAND,GAS STATION,true,false,0932,009,16,61,26,
,0810,THEFT,OVER $500,PARKING LOT/GARAGE(NON.RESID.),false,false,1434,014
,DECEPTIVE PRACTICE,"THEFT BY LESSEE,MOTOR VEH",AIRPORT VENDING ESTABLISH
486,BATTERY,DOMESTIC BATTERY SIMPLE,STREET,false,true,0235,002,5,41,08B,1
```



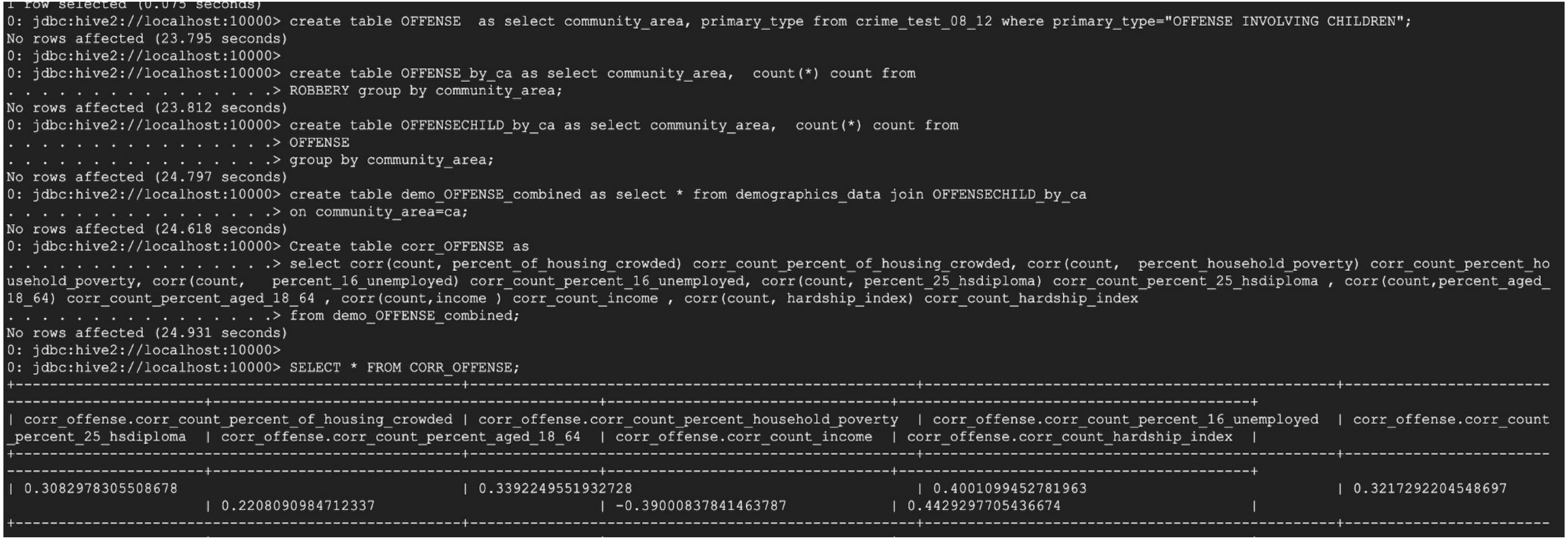
Location
(41.815117282, -87.669999562)

Had to use python scripts to replace commas with semicolons and remove double quotes before data could be cleaned further using MapReduce

Challenge

We had to find correlation between 7 socioeconomic variables with 31 types of crime incident, and then store them into tables for Tableau visualization. Running single “CORR(field1, field2)” meant running 217 queries.

Came up with a set of 4 queries ran for each type of crime, to get all the values in different tables.

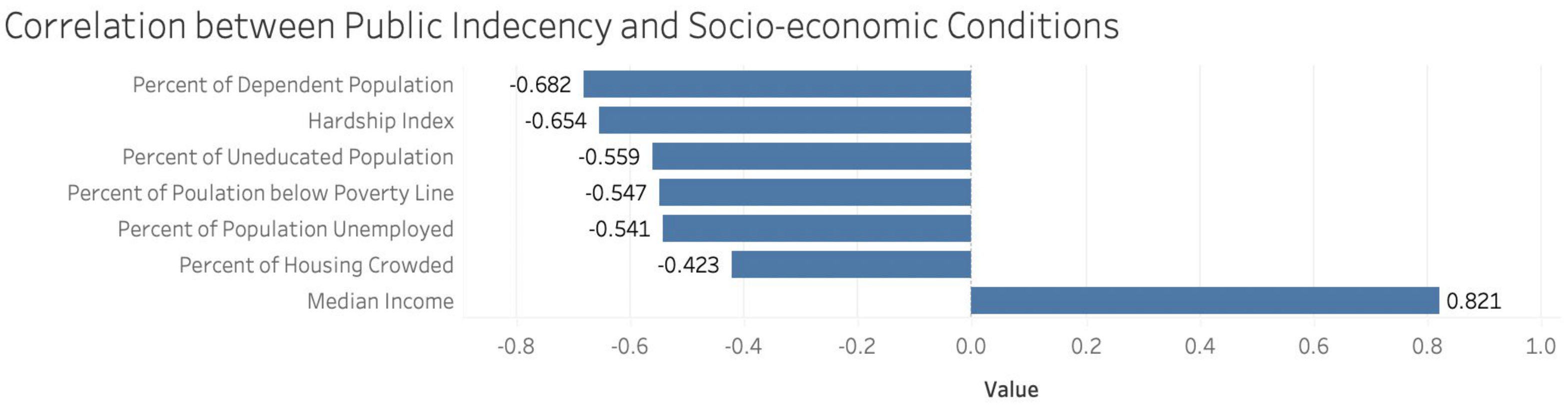


```
1 row selected (0.075 seconds)
0: jdbc:hive2://localhost:10000> create table OFFENSE  as select community_area, primary_type from crime_test_08_12 where primary_type="OFFENSE INVOLVING CHILDREN";
No rows affected (23.795 seconds)
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> create table OFFENSE_by_ca as select community_area,  count(*) count from
. . . . . . . . . . . . . . . . .> ROBBERY group by community_area;
No rows affected (23.812 seconds)
0: jdbc:hive2://localhost:10000> create table OFFENSECHILD_by_ca as select community_area,  count(*) count from
. . . . . . . . . . . . . . . . .> OFFENSE
. . . . . . . . . . . . . . . . .> group by community_area;
No rows affected (24.797 seconds)
0: jdbc:hive2://localhost:10000> create table demo_OFFENSE_combined as select * from demographics_data join OFFENSECHILD_by_ca
. . . . . . . . . . . . . . . . .> on community_area=ca;
No rows affected (24.618 seconds)
0: jdbc:hive2://localhost:10000> Create table corr_OFFENSE as
. . . . . . . . . . . . . . . . .> select corr(count, percent_of_housing_crowded) corr_count_percent_of_housing_crowded, corr(count,  percent_household_poverty) corr_count_percent_ho
usehold_poverty, corr(count,   percent_16_unemployed) corr_count_percent_16_unemployed, corr(count, percent_25_hsdiploma) corr_count_percent_25_hsdiploma , corr(count,percent_aged_
18_64) corr_count_percent_aged_18_64 , corr(count,income ) corr_count_income , corr(count, hardship_index) corr_count_hardship_index
. . . . . . . . . . . . . . . . .> from demo_OFFENSE_combined;
No rows affected (24.931 seconds)
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> SELECT * FROM CORR_OFFENSE;
+----------------------------------------------------+----------------------------------------------------+------------------------------------------------+------------------------
-----------------------+---------------------------------------------+---------------------------------+-----------------------------------------+
| corr_offense.corr_count_percent_of_housing_crowded | corr_offense.corr_count_percent_household_poverty  | corr_offense.corr_count_percent_16_unemployed  | corr_offense.corr_count
_percent_25_hsdiploma  | corr_offense.corr_count_percent_aged_18_64  | corr_offense.corr_count_income  | corr_offense.corr_count_hardship_index  |
+----------------------------------------------------+----------------------------------------------------+------------------------------------------------+------------------------
-----------------------+---------------------------------------------+---------------------------------+-----------------------------------------+
| 0.3082978305508678                                 | 0.3392249551932728                                 | 0.4001099452781963                             | 0.3217292204548697
                       | 0.2208090984712337                          | -0.39000837841463787            | 0.4429297705436674                      |
+----------------------------------------------------+----------------------------------------------------+------------------------------------------------+------------------------
```

Results

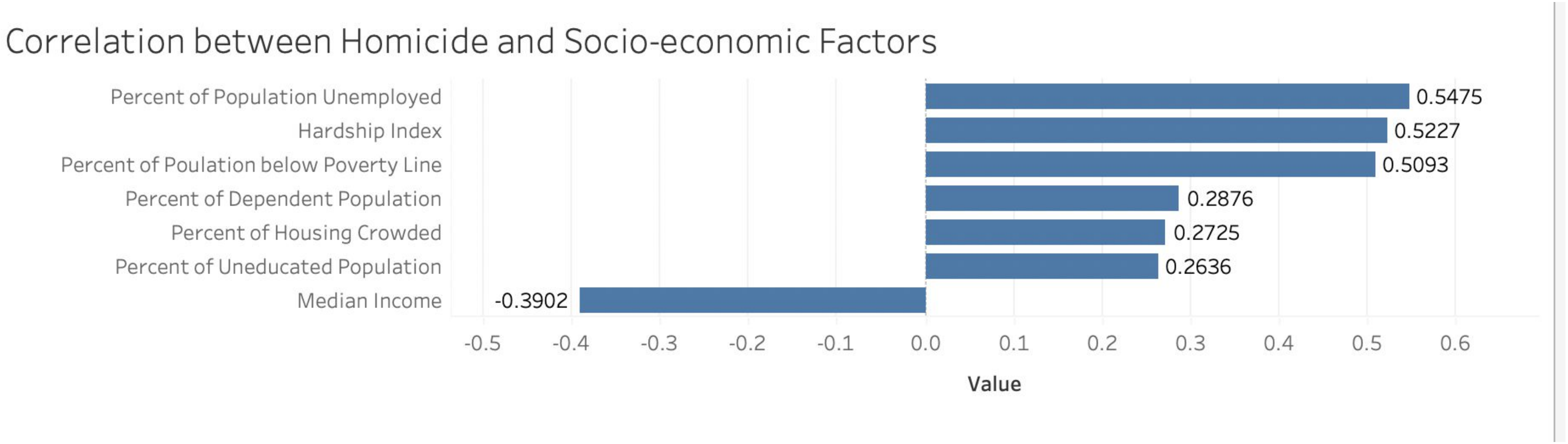
We found that Crime Incidents that involved Public Indecency had a high correlation with the socioeconomic factors.

Areas with better conditions had higher incidents of indecency (only income has a positive correlation, rest have a negative correlation).



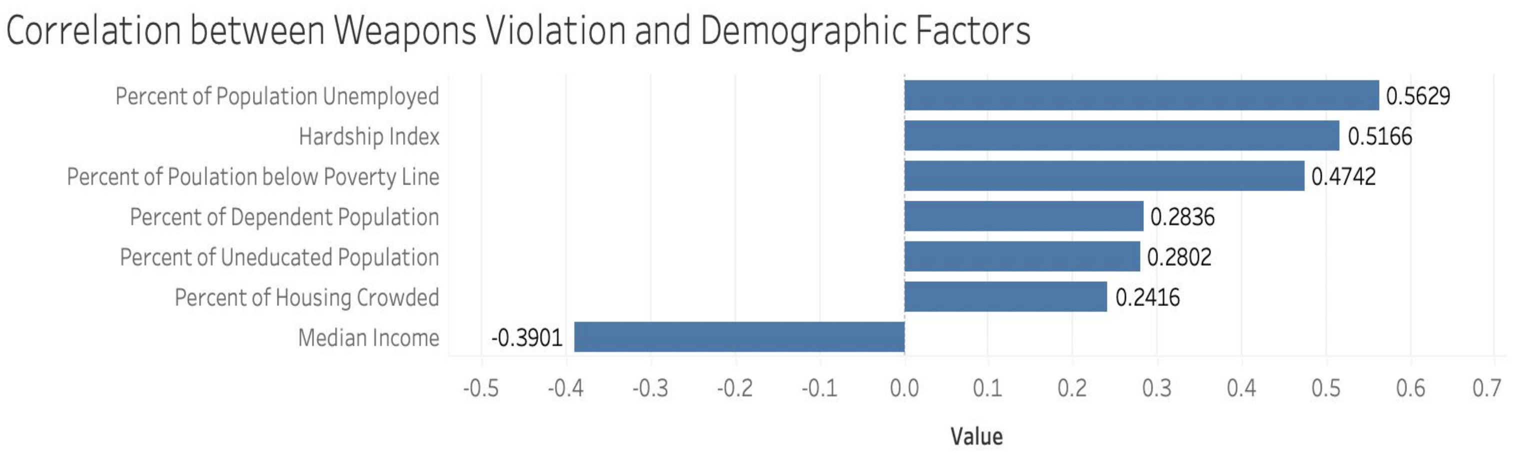
Results

We found that Homicides were positively correlated to poverty, unemployment and overall hardship index.

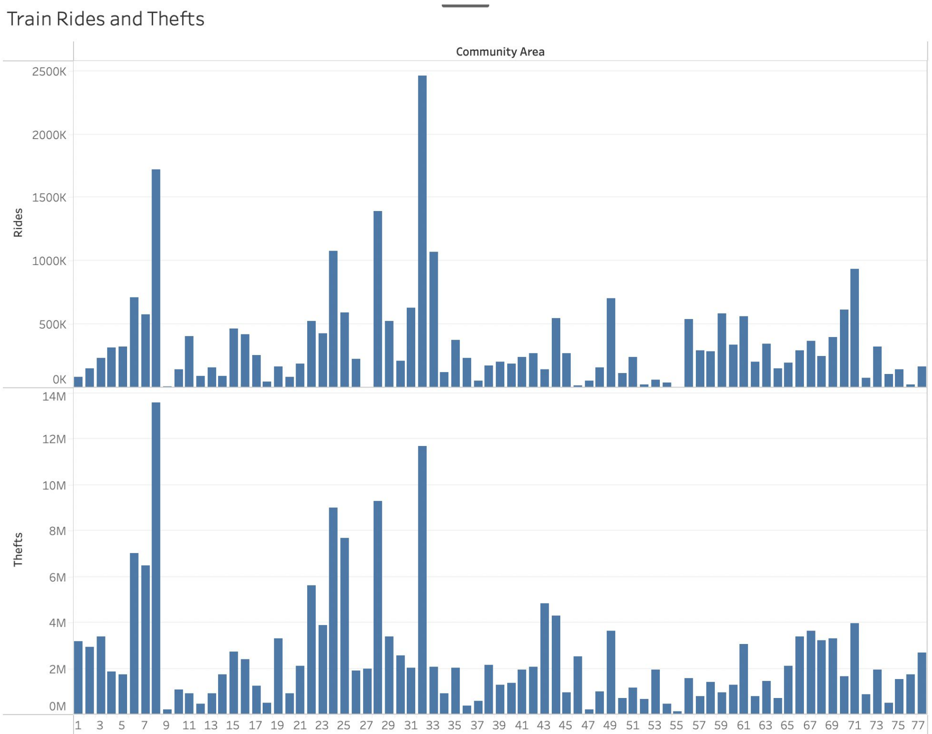


Results

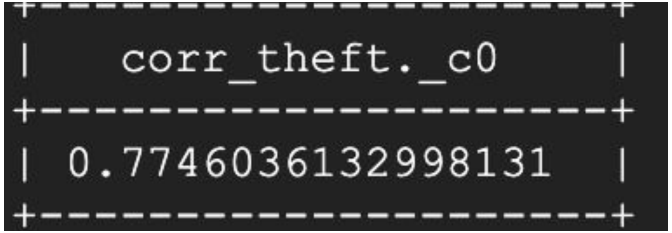
We found that Weapons Violation were positively correlated to poverty, unemployment and overall hardship index.



Results



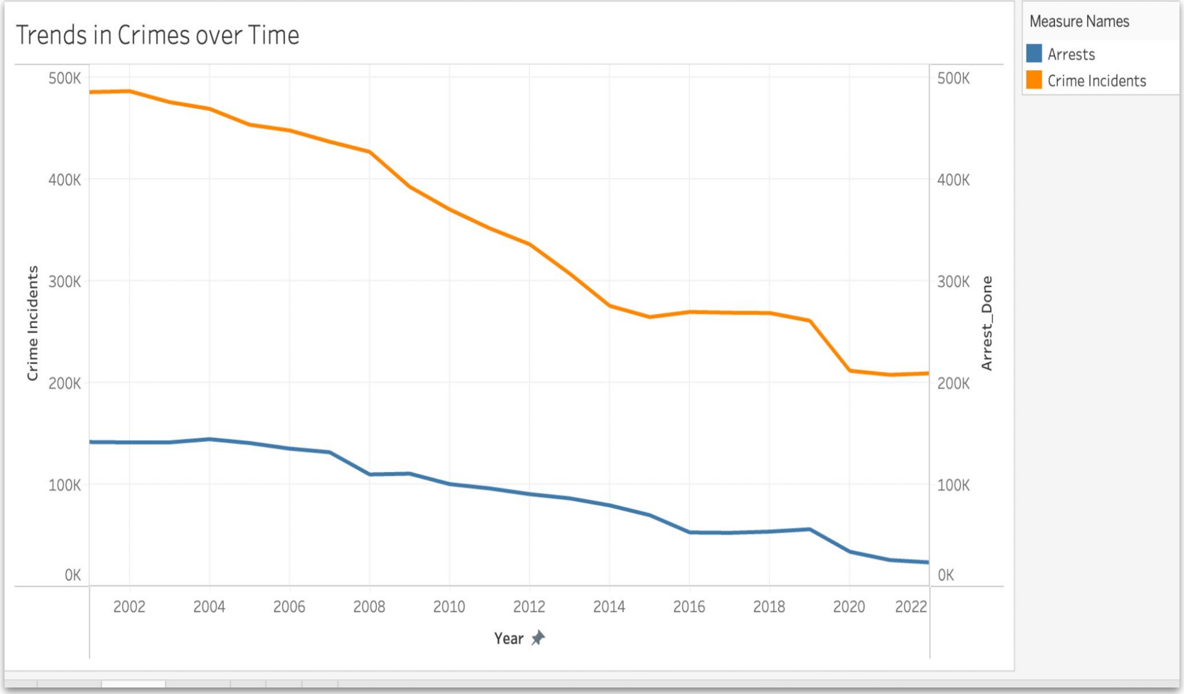
We found that number of train rides is positively correlated (0.77) with the number of thefts in Chicago.



```
+--------------------+
|   corr_theft._c0   |
+--------------------+
| 0.774603613299813 1 |
+--------------------+
```

Results

Overall, the crime incidents have reduced since 2001. Arrest rates follow a similar trend.



Obstacles

- MapReduce for some datasets took hours to finish running.
- Hive does not support all SQL functions.
- Tableau visualizations involved some learning curve.
- Unavailability of data for all years.

Summary

- Crime is highly correlated to Median Income, Hardship Index, and level of education among other expected and unexpected factors.
- Governmental policies should seek to find ways to reduce positively correlated factors.
- Governmental policies should seek to find ways to increase the negatively correlated factors.
- Further work can be done to find multi-layered correlations that will give decision makers a better understanding of crime in Chicago and how to prevent it.

References & Acknowledgments

- 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.
- https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF
- http://hadooptutorial.info/hive-aggregate-functions/
- https://help.tableau.com/current/pro/desktop/en-us/examples_hortonworkshadoop.htm

Thank you!