

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

CRIMINALITY & DEMOGRAPHICS: A CHICAGO CASE

MSc in Business Analytics

Data Management & Business Intelligence

Fall Quarter 2021-2022

Professor: Damianos Chatziantoniou

Students: Argirios Taklakoglou 2822114

Michail Kalligas 2822103

Panagiotis Lolos 2822105

Athens, 2021

Table of Contents

1. Introduction	3
2. About the Datasets	3
2.1. Crimes 2001-to-Present	3
2.2. Census data 2008-2012.....	4
2.3. Community Area 2000 & 2010 Census Population Comparisons.....	4
3. On Data Cleaning	5
4. On ETL Process and Cube.....	5
4.1. Extract & Transform.....	5
4.2. Loading the Database	6
4.3. Process and Deploy the Cube	10
4.4. Browsing the cube and checking values	11
5. Power BI and Visual Reports.....	14
6. Conclusions	21
7. Sources & Tools.....	23

1. Introduction

We assume that we are an insurance startup located in Chicago with plans of expanding our products to life & wealth insurance packages. In order for the pricing, risk and budget to be properly calculated, it is necessary to study not only crime rates but any economic factor that may influence them significantly. Therefore, we attempt to combine, analyze and visualize two datasets taken directly from [Chicago Data Portal](#). After cleaning the datasets, we use traditional ETL procedures to create a database holding all useful information, then create a cube with OLAP services which we use in Power BI to extract our conclusions regarding which factors should be taken into account on our future products and deals, especially when screening new customers for custom offers. Some of these factors include counts for specific crime categories, crime coordinates in association with the community area that they indicate, economic measures for said community areas and the overall crime rates during the years 2008-2012.

2. About the Datasets

The purposes of our study initially revolved around finding correlations between criminality rates and specific demographic features. The city of Chicago is the third most populous city in the USA (Wikipedia, 2021), and has an infamous track record on having the highest crime ratios of all American cities, thus providing an excellent source of information and crime statistics. For the crime metrics, we managed to find the dataset [Crimes 2001-to-Present](#) in the *Public Safety* section of the official Chicago Data Portal. In the same portal, we were also able to find in the *Health & Human Services* section the dataset [Census Data 2008-2012](#) on selected socioeconomic indicators from the 2008-2012 census. Finally, it was essential for the study to incorporate in some way each community area's population, which was conveniently found in a [2010 population census .xls](#) file in the official city of Chicago page.

2.1. Crimes 2001-to-Present

The dataset consists of over 7.45 million entries on reported crimes as they are extracted from the CPD's (Chicago Police Department) CLEAR (Citizen Law Enforcement Analysis and Reporting) system. It has 22 columns, some of which will not be explained as they were not used in the study:

- ID: unique crime record identifier
- Date: exact date of the occurrence
- Block: block location of the incident with redacted address
- IUCR: Illinois Uniform Crime Reporting code, which indicates a type of crime along with its' description
- Primary Type: crime type as per IUCR
- Description: crime type description as per IUCR
- Community Area: indicates the community area where the incident occurred
- FBI Code: indicates the crime code (type) as per FBI's NIBRS (National Incident Based Reporting System)
- Beat: the smallest geographical area defined by the CPD for patrolling beats (cars). In this case, it refers to the beat number that first responded to the crime.
- Year: the year of the occurrence
- Latitude: the estimated earthly latitude on a map of the occurrence

- Longitude: the estimated earthly longitude on a map of the occurrence

The dataset is updated on a daily basis from Tuesday to Sunday and some entries may not be verified, as they are based on preliminary reports to the Police Department and recorded as is. This is extremely helpful in both trying to predict future crime rates at any time and reevaluating factors and features over time whenever it's deemed necessary. At the same time, however, it increases drastically the risk of human error, duplicate creation and a certain degree of volatility when parsing daily data. According to CPD, some errors may be fixed at later dates after further investigation, which is why data is always added at least one week after an occurrence and not earlier.

2.2. Census data 2008-2012

The dataset consists of Chicago's 77 community areas, evaluated with 7 socioeconomic indicators, as extracted by the census that occurred in the years 2008-2012. It has 9 columns:

- Community Area Number: a unique number identifier for each community area
- Community Area Name: the full name of each community area
- Percent of Housing Crowded: an indicator of house (over)crowding meaning the percent of housing units with more than one person per room
- Percent Households Below Poverty: an economic indicator about the percent of households living below the federal poverty threshold
- Percent aged 16+ Unemployed: an unemployment rate for people over 16 years old
- Percent aged 25+ without high school diplomas: an illiteracy indicator for people over 25 years old concerning high school education
- Percent aged under 18 or over 64: percent of population that's considered dependent by the state (dependency indicator)
- Per capita income: an indicator of the average income per community area as calculated by dividing the sum of tract-level aggregate incomes with the total area population
- Hardship Index: the resulting score when taking into account the above indicators for each community area

There was no census data closer to today that offered such distinct information for each community area, so we had to limit our study to the five-year period described by the dataset. This proved useful in that we could avoid the Crimes dataset's volatility of later values by only choosing reports that have been reviewed and verified over time. It also allowed us to reduce the vast number of entries into a solid 25% (1,878,462 entries out of roughly 7,450,000), which was pivotal during the ETL process and the cube processing given our limited equipment. Our goal during analysis was to find correlations between at least the hardship index and criminality rate and if possible, to further break it down into specific factors.

2.3. Community Area 2000 & 2010 Census Population Comparisons

The dataset was a simple xls table consisting of all 77 community areas and their respective populations as counted in the years 2000 and 2010. Additionally, it offered a Difference in population between the two years, as well as a decline/growth percentage. For our data, only the populations recorded in 2010 were needed, which we joined in the demographic dataset using the Community Areas' names as a key.

3. On Data Cleaning

A rigorous amount of data cleaning was necessary over the datasets before using them properly, which is why we created two new datasets (crimes8_12.csv, demogr8_12.csv) using R and RStudio and executed the following:

- Deleted ID columns from both datasets.
- Deleted duplicate rows from crimes.csv.
- Filtered out any crime entry before 2008 and after 2012 in crimes.csv.
- Reconstructed datetime values from crimes.csv in order to store the units of time in the appropriate date and time dimensions of the data warehouse.
- Removed community area “CHICAGO” from census.csv as it does not indicate an actual area.
- Ordered columns in crimes.csv for better readability.
- Manually cleaned up FBI Codes so that errors of misassignments are redacted.
- Changed all NULL/’/NA values to string “Unknown” so that cube dimensions will work properly.

The resulting datasets were a massively smaller crimes8_12.csv (about 35% its’ original size) with only important entries, no duplicates or null values and entries within the five-year period defined by the census, as well as a more refined census8_12.csv so that it is better paired with the former. For the population xls, no data cleaning was required, since we only kept the two important columns (Community Area Name, Population 2010) and there were no irregular values.

4. On ETL Process and Cube

4.1. Extract & Transform

Part of the ETL Process was executed using the same R script that we used for data cleaning. For crimes8_12.csv, we created a new crmsfacts.csv containing the fact table of the crimes.csv and for demogr8-12.csv, we created a new dgrfacts.csv containing the fact table of the census.csv. For crmsfacts.csv we created the dimensions Date, TimeofDay, FBI Code, IUCR code, Primary Type, Crime Description, Location Description, Beat, District, Ward, Community Area, Block. A junk dimension was created to store the dual variables Arrested and Domestic (true or false). For latitude, longitude, x and y coordinates dimensions were not possible (or practical) to create given the excess number in digits and values that the variables may take. All dimensions were connected to crmsfacts.csv via integer keys.

For dgrfacts.csv we only created the dimensions Community Area and Census, where the latter only takes the integer 1 for now. Census dimension can be used in the case of a new census dataset (ex. Census 2) to encapsulate a new period within the crimes dataset in the future so as to acquire newer information. The census dataset only has 77 fixed entries, which make it very hard and impractical to determine new dimensions for any values they may take. In this case, all variables were either distinct rankings, indexes or percentages, none of which offers a good dimension and were used as they are.

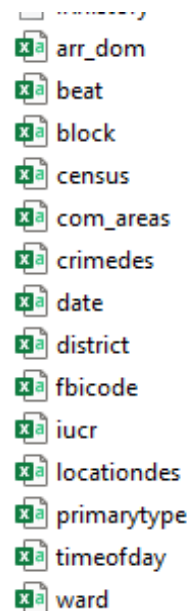


Figure 1
.csv dimensions

As for incorporating the population within the dgrfacts.csv, that was easily done by using the SQL wizard and inner joining the resulting table from the xls file with dgrfacts on Community Name. As we used R for our implementations, for each new dimension a small .csv was created, representing its respective table. The full list of usable dimensions may be seen in *Figure 1*.

4.2. Loading the Database

After having created the dimensions in the form of csv files, creating the database was a relatively easy process using the SSDT's SSIS toolbox. The resulting DB consisted of the two fact tables along with all of their dimension tables as seen in *Figure 2*.

It should be noted that we did not use the original or even the cleaned up csv files for the database load, since we had already created the necessary dimensions and fact tables for the cube during the cleanup process with R.

Pulling up some indicative rows from each table yielded the results below:

- arr_dom (dimension):

It contains every possible binary combination of values for the variables *arrest* and *domestic*. It is essentially a junk dimension.

	arr_domID	arrest	domestic
1	1	0	0
2	2	0	1
3	3	1	0
4	4	1	1

- Beat (dimension):

It contains all 303 geographical areas where a beat car is assigned by the CPD. Each area has a unique beatID and a three digit number corresponding to it.

	beatID	beat
1	1	111
2	2	112
3	3	113
4	4	114
5	5	121
6	6	122
7	7	123

- Block (dimension):

It contains all 32,596 building blocks in Chicago. Each is assigned a unique blockID and an address. As a data protection measure, the CPD omits the actual address/flat number of the location of a crime and instead replaces it with the general block area where it happened, thus making a block the smallest possible geographical area to pinpoint for each crime.

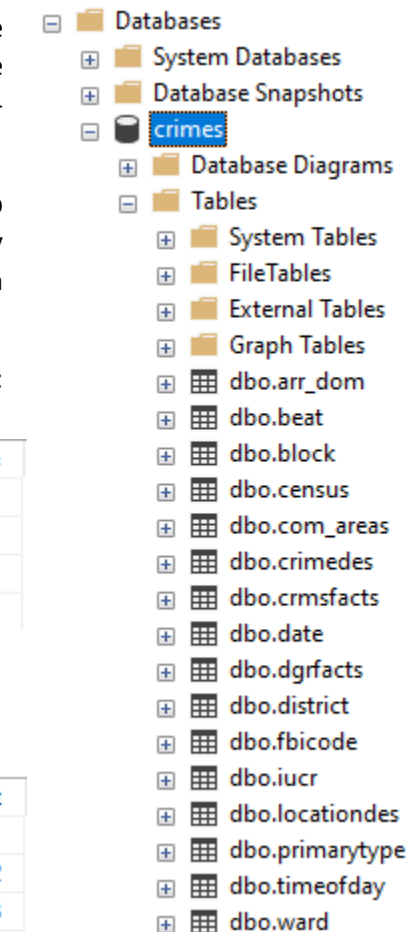


Figure 2: Database as seen in the SSMS Explorer

	blockID	block
1	1	0000X E 8TH ST
2	2	0000X E 9TH ST
3	3	0000X E 11TH ST
4	4	0000X E 13TH ST
5	5	0000X E 14TH PL
6	6	0000X E 14TH ST

- Com_areas (dimension):

It contains all 77 community areas in Chicago. Each is assigned a unique incremental ID, which is duplicated from the com_areas column (their official number) as a connector to both datasets. Each area also has its own string name, with the exception of area 78, which was initially “CHICAGO”, indicating the general area of Chicago. We decided to change it to “Unknown” as it signifies that the area of the crime is unspecified.

	com_areasID	com_areas	com_areas_nms
1	1	1	Rogers Park
2	2	2	West Ridge
3	3	3	Uptown
4	4	4	Lincoln Square
5	5	5	North Center

- Crimesdes (dimension):

It contains all 369 Crime Descriptions according to CPD. Each is assigned a unique incremental ID, so that every possible crime code is properly described.

	crimesdesID	crimesdes
1	1	\$500 AND UNDER
2	2	ABUSE/NEGLECT: CARE FACILITY
3	3	AGG CRIM SEX ABUSE FAM MEMBER
4	4	AGG CRIMINAL SEXUAL ABUSE
5	5	AGG PO HANDS ETC SERIOUS INJ
6	6	AGG PO HANDS NO/MIN INJURY

- Date (dimension):

It is a simple date dimension, containing all possible dates when crimes happened in Chicago during the period 2008-2012. It contains an incremental ID for each date (starting from the oldest possible date to the newest), the name of each day and three integers for day, month and year.

	dateID	wholedate	nameofday	day	month	year
1	1	2008-01-01	Tuesday	1	1	2008
2	2	2008-01-02	Wednesday	2	1	2008
3	3	2008-01-03	Thursday	3	1	2008
4	4	2008-01-04	Friday	4	1	2008
5	5	2008-01-05	Saturday	5	1	2008
6	6	2008-01-06	Sunday	6	1	2008

- District (dimension):

It contains a small list of all 23 police districts in Chicago. Each is assigned an incremental ID which sometimes is the same as its own numbering. For example, district 31 is given the ID 23, since districts 26-30 do not exist. Same as with the Block dimension, ID 24 means “Unknown” for crimes of unknown location.

	districtID	district
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5

- Fbicode (dimension):

It contains all 25 codes for catalogued federal offenses, as encoded by the FBI. It has a similar function to crimedes. Each crime is assigned an ID, an incremental string number, a type and its official definition. Code 26 is for miscellaneous non-index offenses.

	fbicodeID	fbicode	crimetype	definition
1	1	01A	Homicide 1st & 2nd Degree	The killing of one human
2	2	01B	Involuntary Manslaughter	The killing of another per
3	3	02	Criminal Sexual Assault	Any sexual act directed a
4	4	03	Robbery	The taking or attempting
5	5	04A	Aggravated Assault	An unlawful attack by on
6	6	04B	Aggravated Battery	An unlawful attack by on

- Iucr (dimension):

It contains all 369 crime codes as used by the CPD. It is directly linked to Crimedes and Primarytype dimensions. Each crime code is assigned a unique incremental ID, which may then be used to find the crime type from Primarytype as well as the description from Crimedes.

	iucrID	iucr
1	1	031A
2	2	031B
3	3	033A
4	4	033B
5	5	041A
6	6	041B

- Locationdes (dimension):

It contains all 159 location descriptions referenced in the dataset. Code 1 refers to any crime lacking sufficient description.

	locationdesID	locationdes
1	1	Unknown
2	2	ABANDONED BUILDING
3	3	AIRCRAFT
4	4	AIRPORT BUILDING NON-TERMINAL - N
5	5	AIRPORT BUILDING NON-TERMINAL - SI

- Primarytype (dimension):

It contains all 32 official distinct crime categories, under which fall the IUCR codes and descriptions. Each category is assigned a unique incremental ID.

	primarytypeID	primarytype
1	1	ARSON
2	2	ASSAULT
3	3	BATTERY
4	4	BURGLARY
5	5	CRIM SEXUAL ASSAULT
6	6	CRIMINAL DAMAGE
7	7	CRIMINAL SEXUAL ASSAULT
8	8	CRIMINAL TRESPASS

- Timeofday (dimension):

It contains all minute timestamps of a single day, starting at 00:00:00 and ending at 23:59:00. Each timestamp is assigned a unique incremental ID up to 1440. Timestamps are used to pinpoint the exact moment of a crime occurrence.

	timeofdayID	timeofday
1	1	00:00:00.0000000
2	2	00:01:00.0000000
3	3	00:02:00.0000000
4	4	00:03:00.0000000
5	5	00:04:00.0000000
6	6	00:05:00.0000000

- Ward (dimension):

It contains all 50 City Council Districts of Chicago. Each (called a Ward) is assigned a unique number from 1 to 50, representing a large area much like community areas but delimited for different purposes. Just like with other area dimensions, number 51 is assigned to “Unknown” locations.

	wardID	ward
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5

- Crmsfacts (fact table):

	crimeID	casenumber	updatedon	dateID	timeofdayID	fbicodeID	iucrID	primarytypeID	crimesesID	locationdesID	arr_domID	beatID	districtID	wardID	com_areasID	blockID	x_coord	y_coord	latitude	longitude
1	1	HP115238	2018-02-28 ...	9	1106	8	100	31	1	143	1	187	14	37	25	22142	1139890	1904803	41.89...	-87.76...
2	2	HP115211	2018-02-28 ...	9	1110	4	1	28	57	24	1	288	21	48	77	23958	1168537	1940249	41.99...	-87.65...
3	3	HP116583	2018-02-28 ...	9	1111	4	34	28	320	146	1	115	9	16	61	21683	1167240	1869424	41.79...	-87.66...
4	4	HP115355	2018-02-28 ...	9	1111	11	50	3	312	17	3	209	16	40	13	14998	1153289	1937021	41.98...	-87.71...
5	5	HP137242	2018-02-28 ...	9	1111	7	95	4	163	120	1	40	4	8	45	28425	1187266	1851321	41.74...	-87.58...
6	6	HP115281	2018-02-28 ...	9	1111	11	62	3	141	19	2	273	18	46	3	4856	1169262	1929404	41.96...	-87.65...
7	7	HP115929	2018-02-10 ...	9	1111	17	166	6	339	119	1	56	5	34	53	31912	1176680	1828327	41.68...	-87.62...
8	8	HP115471	2018-02-10 ...	9	1111	11	50	3	312	146	1	57	5	9	53	32149	1177103	1825158	41.67...	-87.62...
9	9	HP115247	2018-02-10 ...	9	1111	11	62	3	141	119	1	122	10	24	29	16418	1151846	1892457	41.86...	-87.71...
10	10	HP115367	2018-02-10 ...	9	1111	8	99	31	247	63	1	301	22	31	19	8950	1144086	1912565	41.91...	-87.74...

It contains every necessary information for every reported crime, either in the form of degenerate dimensions (like the unique crimeID, the case number, the update log timestamp and location coordinates) or by using IDs referencing foreign keys to any of the above dimensions. The end result is a fairly lightweight but dense fact table to use in our OLAP implementation.

- Census (dimension):

It only contains one entry at present but was created with the future in mind. Censuses happen approximately every 8 to 10 years in Chicago, meaning that one could be ongoing as of the writing of this study. The information extracted from them could be invaluable, thus the need for a dimension indicating when the census was carried out (from which year to which year), along with its unique ID number.

- Dgrfacts (fact table):

	com_areasID	censusID	population	prc_housingcrowded	prc_households_belowpoverty	prc_16plus_unemployed	prc_25plus_nohs	prc_dependent	income	hardship_index
1	1	1	54991	7.7	23.6	8.7	18.2	27.5	23939	39
2	2	1	71942	7.8	17.2	8.8	20.8	38.5	23040	46
3	3	1	56362	3.8	24	8.9	11.8	22.2	35787	20
4	4	1	39493	3.4	10.9	8.2	13.4	25.5	37524	17
5	5	1	31867	0.3	7.5	5.2	4.5	26.2	57123	6
6	6	1	94368	1.1	11.4	4.7	2.6	17	60058	5
7	7	1	64116	0.8	12.3	5.1	3.6	21.5	71551	2
8	8	1	80484	1.9	12.9	7	2.5	22.6	88669	1
9	9	1	11187	1.1	3.3	6.5	7.4	35.3	40959	8

The second fact table of the database contains every socioeconomic indicator presented for the 2008-2012 census. The census was carried out by community areas. Each area is assigned the censusID connecting it to the census dimension, a percent of housing crowded index, a percent of households living below poverty index, a 16+ unemployment rate, a 25+ no high school diploma rate, a percent of dependent people, an average yearly income per capita and a hardship index that encapsulates all of the previous

indexes. A population column was also added manually with each community area's population as recorded in the 2010 census.

4.3. Process and Deploy the Cube

With two dense fact tables and 14 dimensions, we took advantage of both fact tables sharing the same community area IDs to join them via a dimension. The result is the following star schema.

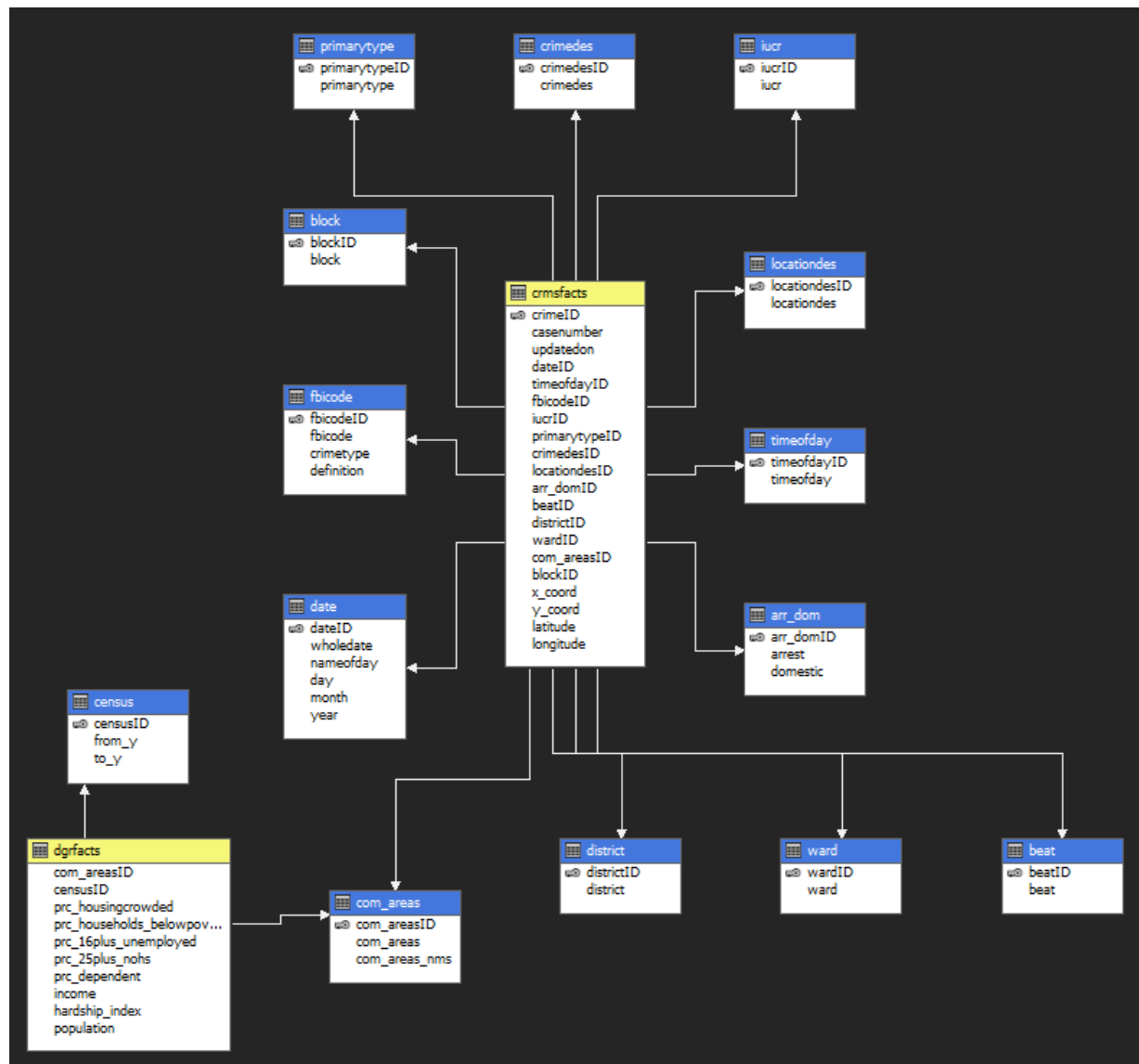
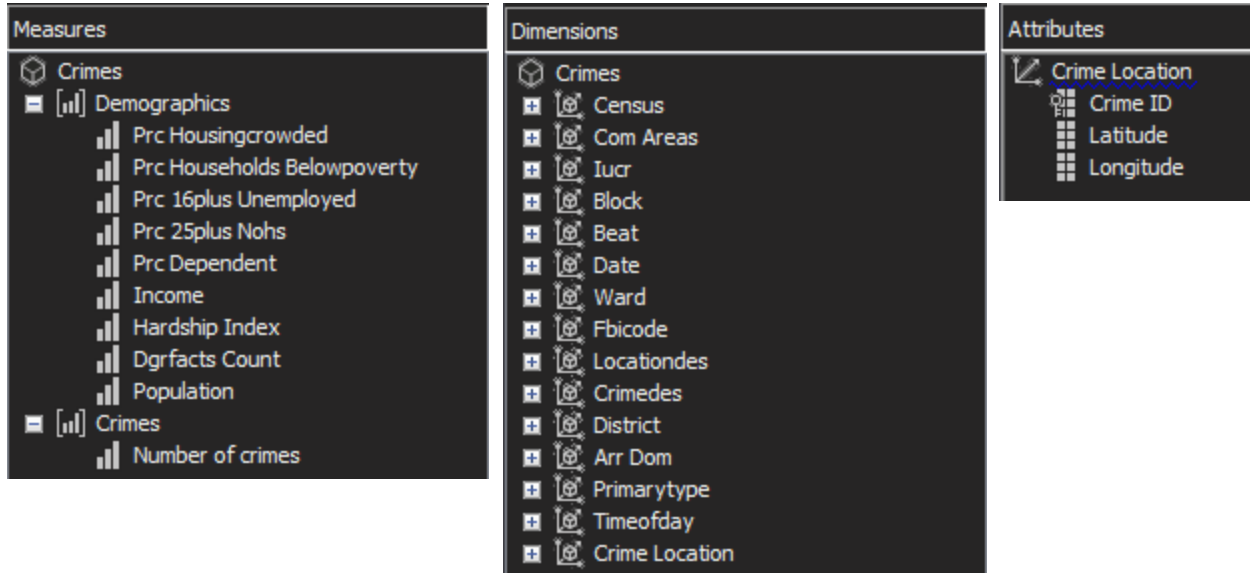


Figure 3: Star schema

The processed cube had 9 measures for the Demographics fact table (as seen below) and only one for the Crimes fact table. We deemed any more metrics for the Crimes fact table unnecessary, as our main focus was to evaluate demographics factors and not study the nature and features of the crimes. There was, however a need for a pivotal degenerate dimension “Crime Location”, which we would use to pinpoint each crime’s location to a certain Community Area.



After processing the schema and deploying the cube, we were almost ready to start analyzing the data using Power BI.

Measure Groups		
Dimensions	Demographics	Crimes
Census	Census ID	
Com Areas	Com Areas ID	Com Areas ID
Iucr		Iucr ID
Block		Block ID
Beat		Beat ID
Date		Date ID
Ward		Ward ID
Fbicode		Fbicode ID
Locationdes		Locationdes ID
Crimedes		Crimedes ID
District		District ID
Arr Dom		Arr Dom ID
Primarytype		Primarytype ID
Timeofday		Timeofday ID
Crime Location		Crime ID

4.4. Browsing the cube and checking values

Before moving on to Power BI, we wanted to make sure that the cube was properly initialized, with all dimensions working as intended. We also wanted to add another metric. To do so, we first needed to inspect the dimensions, conduct some simple experiments with all measure groups and test our metrics. It was a simple task using SSDT's Dimension Usage tab.

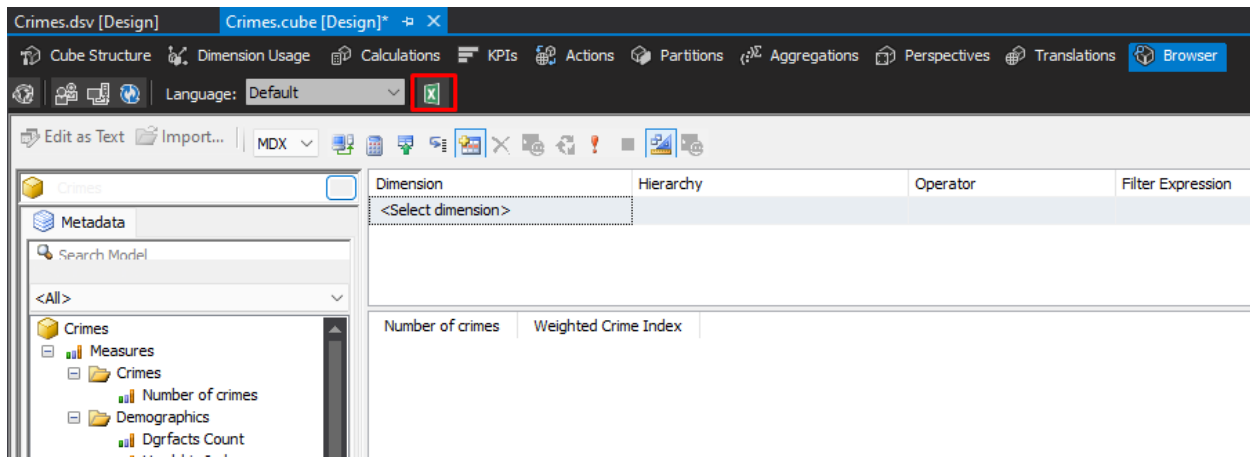


Figure 5 Excel cube browser

	A	B	C
1	Row Labels	Population	Number of crimes
2	Austin	98514	121710
3	South Shore	49767	63364
4	Humboldt park	56323	60261
5	Near North Side	80484	59922
6	West Englewood	35505	57120
7	West Town	81432	55725
8	Auburn Gresham	48743	55028
9	North Lawndale	35912	53119
10	Roseland	44619	52956
11	Englewood	30654	51760
12	Near West Side	54881	50868
13	Chicago Lawn	55628	50812
14	Greater Grand Crossing	32602	47238
15	Loop	29283	40034
16	Logan Square	73595	39936
17	Chatham	31028	39668
18	New City	44377	39485
19	Lake View	94368	38669
20	Belmont Cragin	78743	37232
21	South Chicago	31198	35724
22	East Garfield Park	20567	34671
23	West Garfield Park	18001	33171
24	South Lawndale	79288	31793
25	Woodlawn	25983	31639
26	West Pullman	29651	31550
27	Lincoln Park	64116	29037
28	Rogers Park	54991	28095
29	Uptown	56362	26024
30	Grand Boulevard	21929	24253
31	Portage Park	64124	24226
32	West Ridge	71942	24051
33	Washington Height	26493	22426
34	Irving Park	53359	22362
35	Washington Park	11717	20010

Figure 4: Testing the cube by pivoting

Browsing the cube using SSDT's Excel extension was another way to check all dimensions and values. We used the Pivot Table Fields to navigate the cube and ensure that different dimensions could be combined successfully. In *Figure 5* we confirm our success in properly adding a Population to each Community Area and test the crimes counter metric that we created earlier.

We also decided to add an extra measure to the OLAP services, which would be very useful in determining actual criminality rates. The concept is that simple crime counters are insufficient metrics of criminality per community area because of large differences in population. Thus, a better proxy for criminality in an area would be the number of crimes per area adjusted for population. We named this measure *Weighted Crime index (WCI)*. For a community area i :

$$WCI_i = \text{Number of Crimes}_i * \frac{\text{Average Population}}{\text{Population}_i}$$

The result would be an index that shows the number of crimes in an area, assuming it has the average population of a community area in Chicago. To do that, we first made sure that we could execute SQL aggregates over the population column by attempting to calculate the total and average population of all areas in the same manner we were able to calculate total and average crime numbers for all areas.

	B	C	D	E	F	G
	Population	Number of crimes				
		880	35007.77			
	2695598	1878462				

Next, we used the Calculations tab in the cube design wizard to create a measure script containing the expression as seen in *Figure 6*. Finally, we associated the new measure with the demographics fact table to maintain logical consistency. The end result of using the measure can be seen in *Figure 7* where we

Crimes.dsv [Design] | Crimes.cube [Design]*

Cube Structure | Dimension Usage | **Calculations** | KPIs | Actions | Partitions | Aggregations | Perspectives | Translations | Browser

Script Organizer

- 1. CALCULATE
- 2. [Weighted Crime Index]

Name: [Weighted Crime Index]

Parent Properties

Parent hierarchy: Measures

Parent member: [Change]

Expression

$([Measures].[Number\ of\ crimes]/[Measures].[Population])*35007.77$

Additional Properties

Format string: []

Visible: True

Non-empty behavior: []

Associated measure group: Demographics

Display folder: []

Color Expressions: []

Calculation Tools

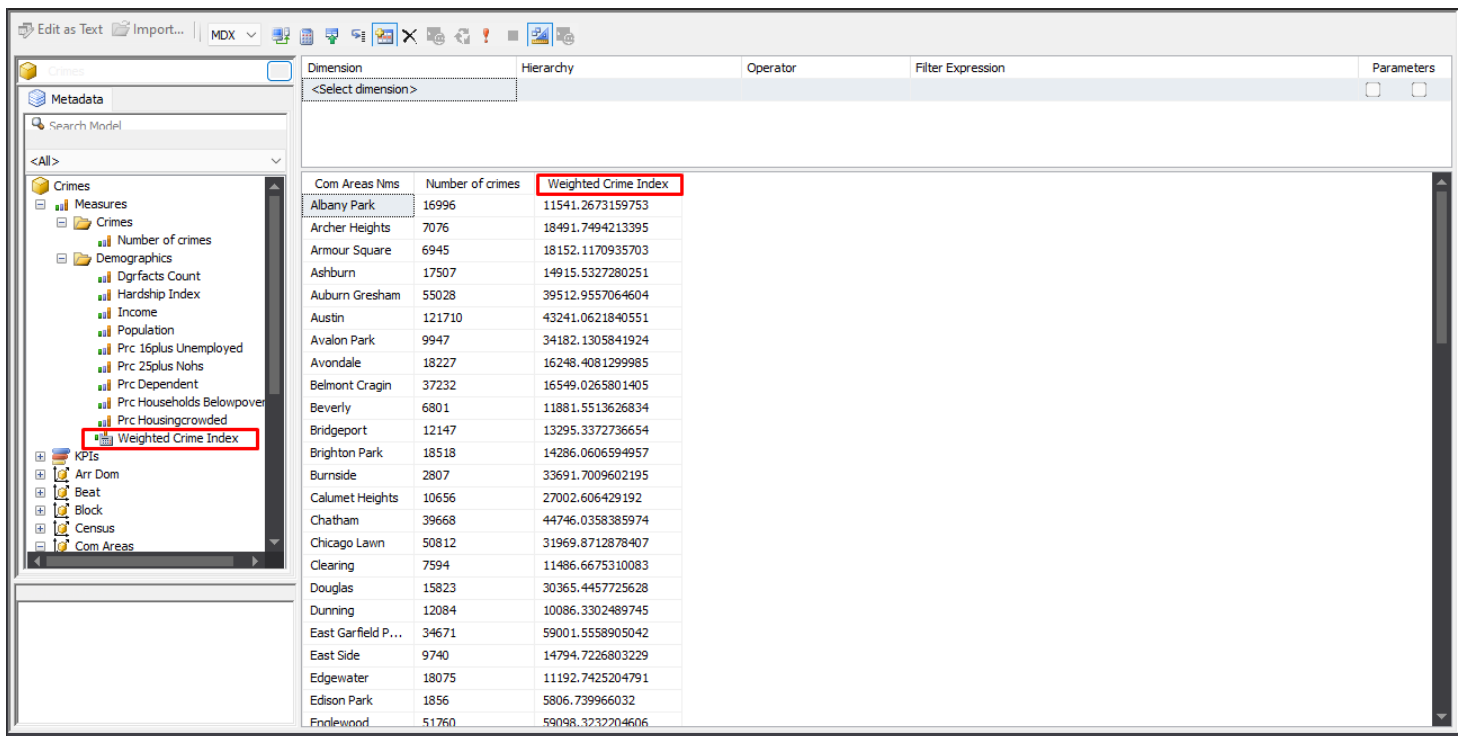
Metadata | Functions | Templates

Search Model

Measure Group: []

Figure 6 Weighted Crime Index Calculation

executed a simple SQL query selecting all Community Areas, their Crime counters and their Weighted Crime indices.



Com Areas Nms	Number of crimes	Weighted Crime Index
Albany Park	16996	11541.2673159753
Archer Heights	7076	18491.7494213395
Armour Square	6945	18152.1170935703
Ashburn	17507	14915.5327280251
Auburn Gresham	55028	39512.9557064604
Austin	121710	43241.0621840551
Avalon Park	9947	34182.1305841924
Avondale	18227	16248.4081299885
Belmont Cragin	37232	16549.0265801405
Beverly	6801	11881.5513626834
Bridgeport	12147	13295.3372736654
Brighton Park	18518	14286.0606594957
Burnside	2807	33691.7009602195
Calumet Heights	10656	27002.606429192
Chatham	39668	44746.0358385974
Chicago Lawn	50812	31969.8712878407
Clearing	7594	11486.6675310083
Douglas	15823	30365.4457725628
Dunning	12084	10086.3302489745
East Garfield P...	34671	59001.5558905042
East Side	9740	14794.7226803229
Edgewater	18075	11192.7425204791
Edison Park	1856	5806.739966032
Englewood	51760	59098.3232204606

Figure 7 Weighted Crime Index in action

5. Power BI and Visual Reports

After all was said and done, all that was left was to visualize the cube and test our hypotheses one by one. The rest of this chapter will be formatted into dashboards, each with its' own description and analysis. Out of our many visual experiments we decided to keep only the dashboards that helped us reach useful conclusions and information regarding our original goals.

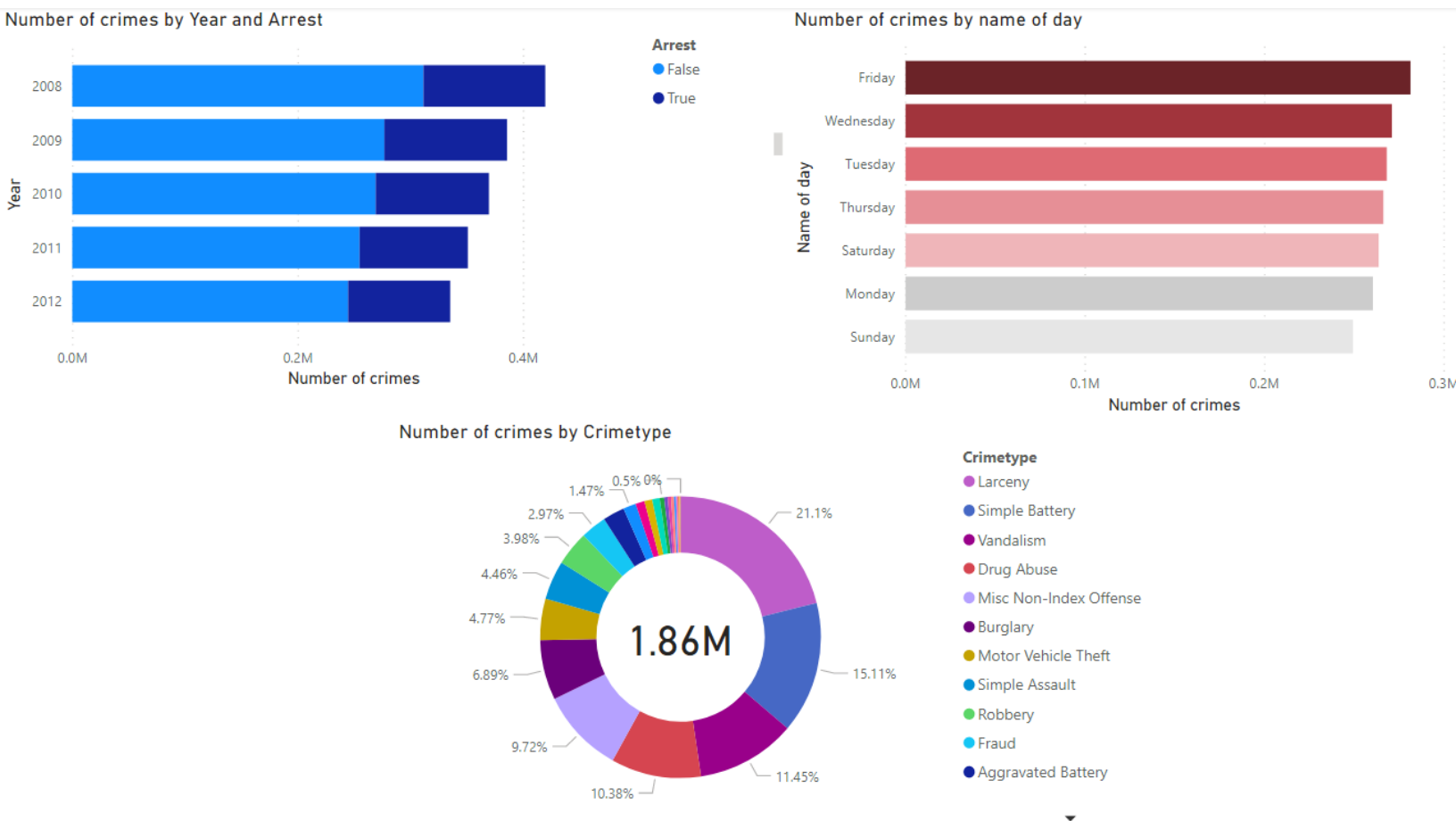
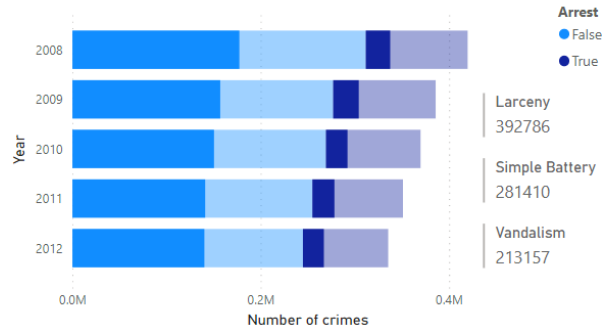


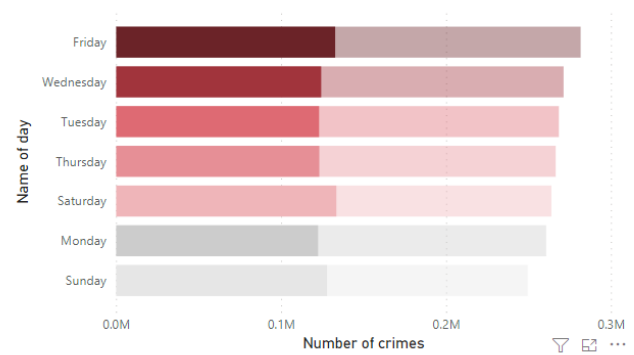
Figure 8 General Information #1

Figure 8 is meant to be an overview of Chicago's criminality. In the upper left histogram, we visualize the total number of crimes committed per year, as well as whether there have been any arrests for them. We find that criminality during that time period is on the decline, even though the percentage of unsolved (without arrests) crimes seems to remain relatively high (approximately 80%). We may thus assume, that the decline in criminality is not a consequence of CPD's contribution but rather one of other factors. In the upper right histogram, we rank weekdays according to the number of crimes committed each day. Although with small differences, Friday seems to be the day when criminals tend to be most active. This could be a useful factor in increasing police patrols and shifts at specific weekdays to maximize effectiveness but is otherwise of no use to us. In the bottom pie chart, we visualize all crime types. We find that over 1.86 million crimes have been committed in the years 2008-2012, 21.1% of which are larcenies, 15.1% of which are simple battery, 11.4% are vandalism, 10.4% are drug abuse and 9.7% are non-index offenses. These 5 categories seem sufficient for us to conduct further research. Larcenies and vandalism are important for property insurance policies, simple assaults are important for health, safety & life insurance packages and drug abuse may result in a variety of insurance breaches. The 5th category, classified as a miscellaneous category also includes any mistyped or non-typed crimes, which makes it especially useful for measuring a more general criminality sample.

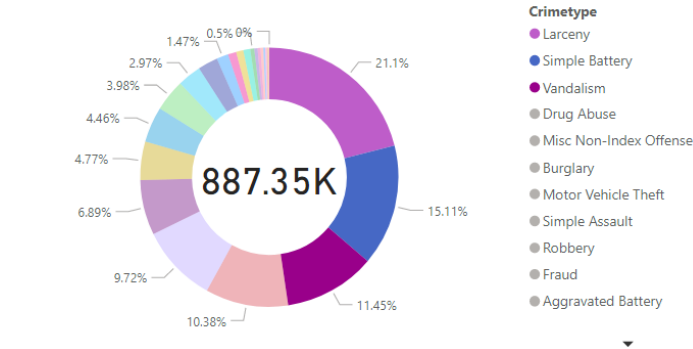
Number of crimes by Year and Arrest



Number of crimes by name of day



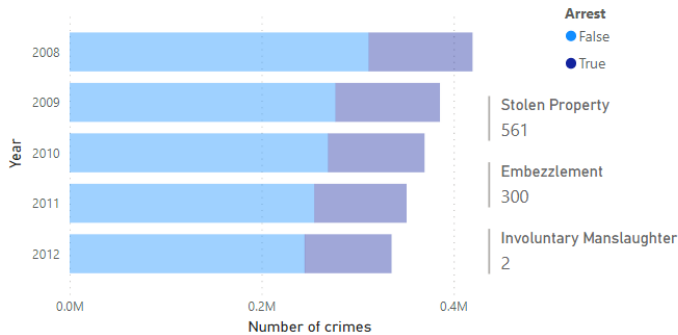
Number of crimes by Crimetype



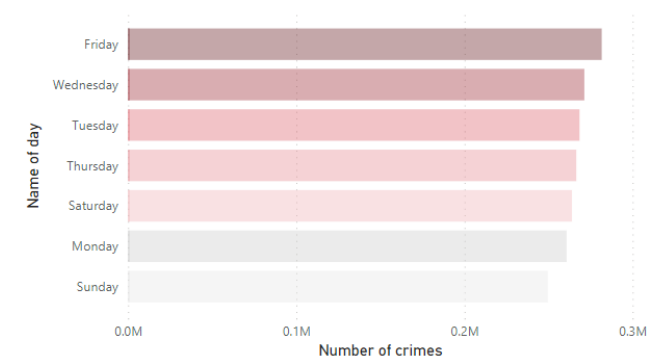
Crimetype	Definition
Simple Battery	A person commits battery if he intentionally or knowingly without legal justification and by any means, (1) causes bodily harm to an individual or (2) makes physical contact of an insulting or provoking nature with an individual.
Larceny	The unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another person.
Vandalism	To willfully or maliciously destroy, damage, deface, or otherwise injure real or personal property without the consent of the owner or the person having custody or control of it.

Figure 9 General Information #2

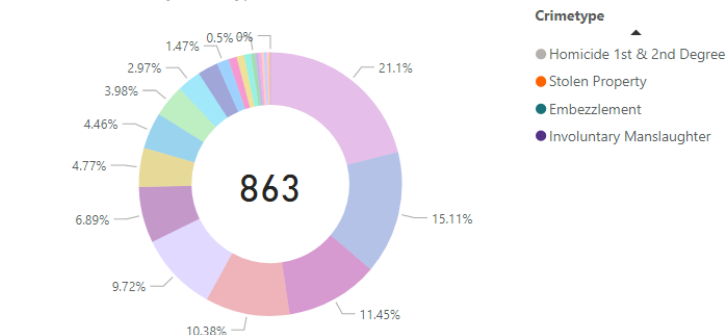
Number of crimes by Year and Arrest



Number of crimes by name of day



Number of crimes by Crimetype



Crimetype	Definition
Stolen Property	Receiving, buying, selling, possessing, concealing, or transporting any property with the knowledge that it has been unlawfully taken, as by Burglary, Embezzlement, Fraud, Larceny, Robbery, etc.
Involuntary Manslaughter	The killing of another person through negligence.
Embezzlement	The unlawful misappropriation by an offender to his/her own use or purpose of money, property, or some other thing of value entrusted to his/her care, custody, or control.

Figure 10 General Information #3

Figures 9 and 10 is but another instance of Figure 9, where we demonstrate the dashboard's ability to be highly interactive. It may be noted that in the bottom right corner, a dictionary has appeared, which we used to further comprehend each of the crime types of significance as we decided previously. In the same way, we were able to determine the crimes with the least occurrences, such as Homicide, Embezzlement, Involuntary Manslaughter and trading of Stolen Property. Since life insurances must take into account homicide and manslaughter rates, we decided that in this case, death by murder has a very small probability of occurring, which actually makes these numbers fairly insignificant when calculating risk.

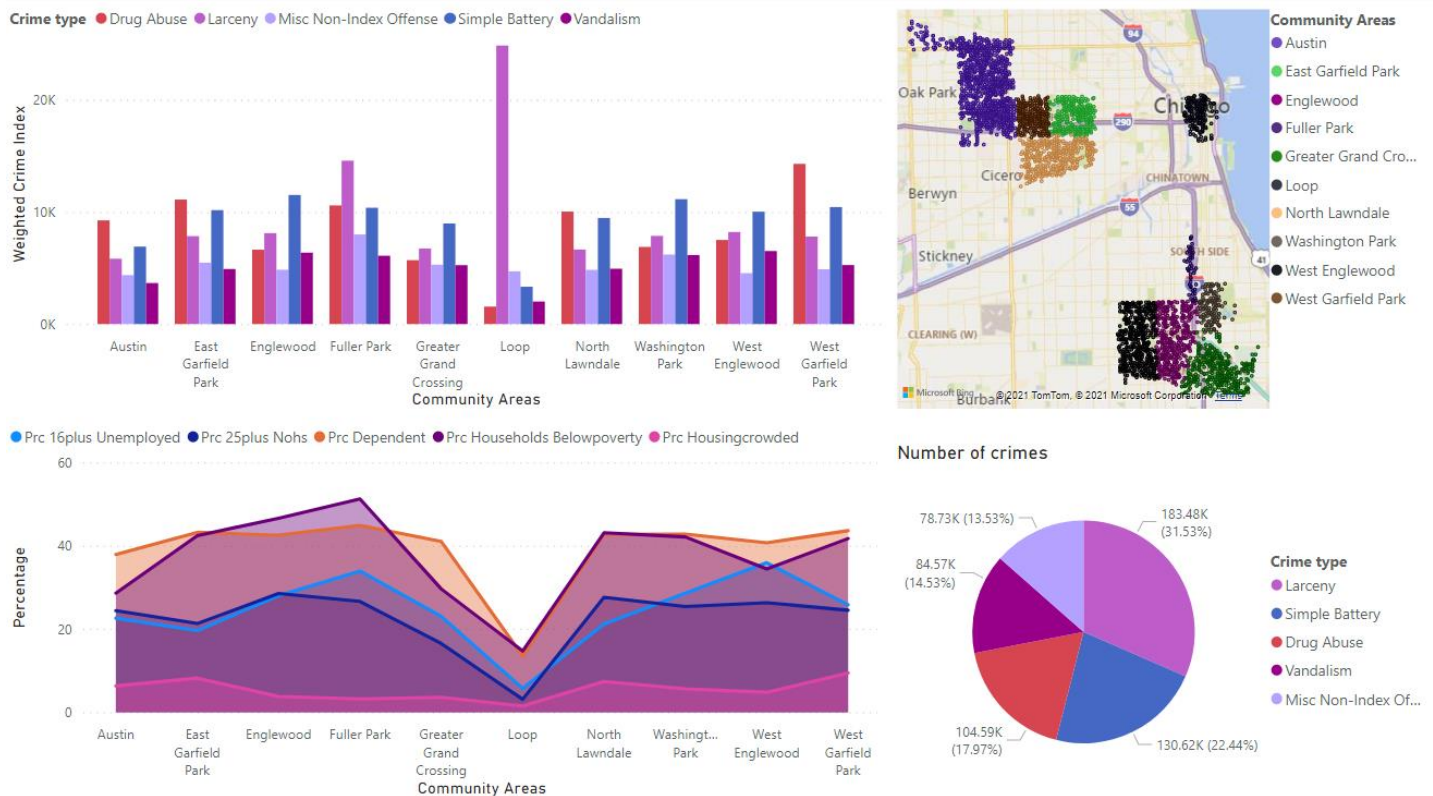


Figure 11 Top 10 areas with high criminality rate

In Figure 11 we have ranked and chosen the ten areas with the highest criminality rate, then filtered through their crimes by focusing on the 5 previously determined most significant categories (Larceny, Simple Battery, Drug Abuse, Vandalism, Misc Non-Index Offense). Looking at the upper right map, we tried to mark territories, where falling a victim of any of those 5 types would be more probable. Sure enough, we noticed two clusters, one around *Fifth City* and one around *Roseland*. To further prove our point, we used the histograms and diagrams in the left side of the dashboard, where we collate each area's socioeconomic indicators (the ones that contribute towards the hardship index with the exception of income) along with their respective WCIs. What we noticed was that areas with extremely bad socioeconomic indicators are prone to battery crimes and drug abuse, whereas areas with much better indicators are plagued by vastly more larcenies. Vandalism and other crimes seem to be somewhat the same no matter the area. Simply put, if a client lives in a lower-class neighborhood, it's more likely he will be assaulted and if he lives in a higher-class neighborhood, it's much more likely he will fall victim to theft. However, we needed to generalize this hypothesis for the entirety of Chicago.

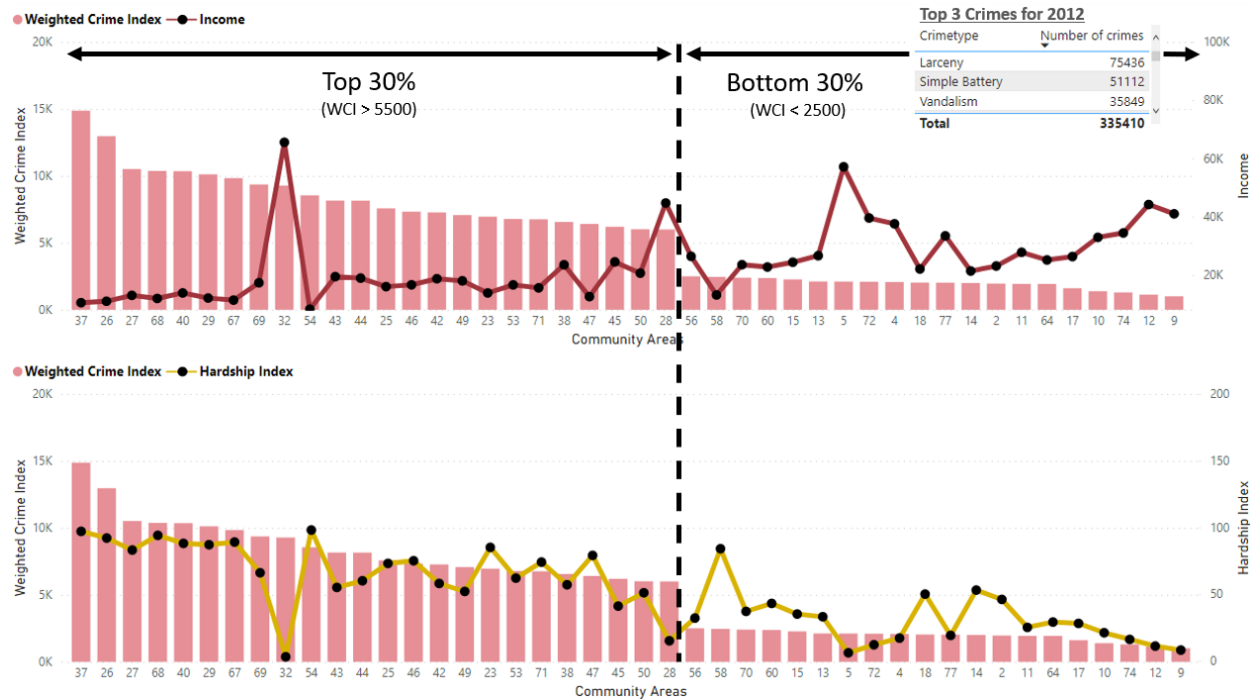


Figure 12 WCI, Income and Hardship Index per community area for 2012

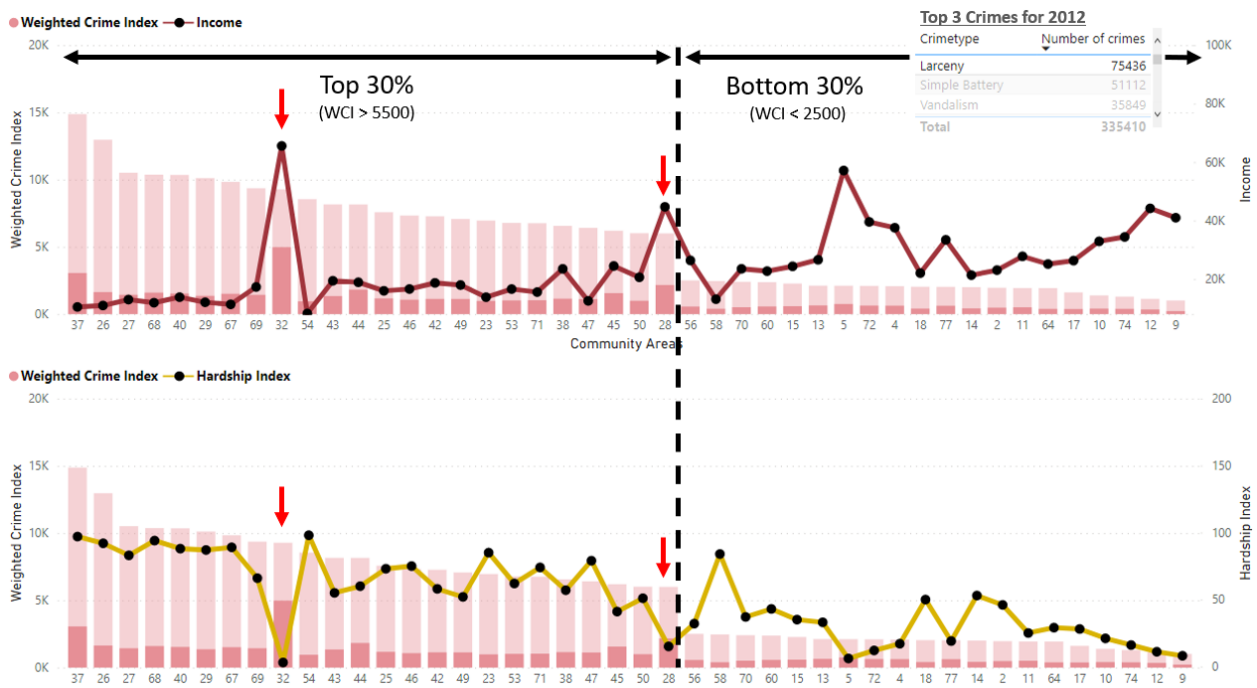


Figure 13 WCI, Income and Hardship Index per community area larceny

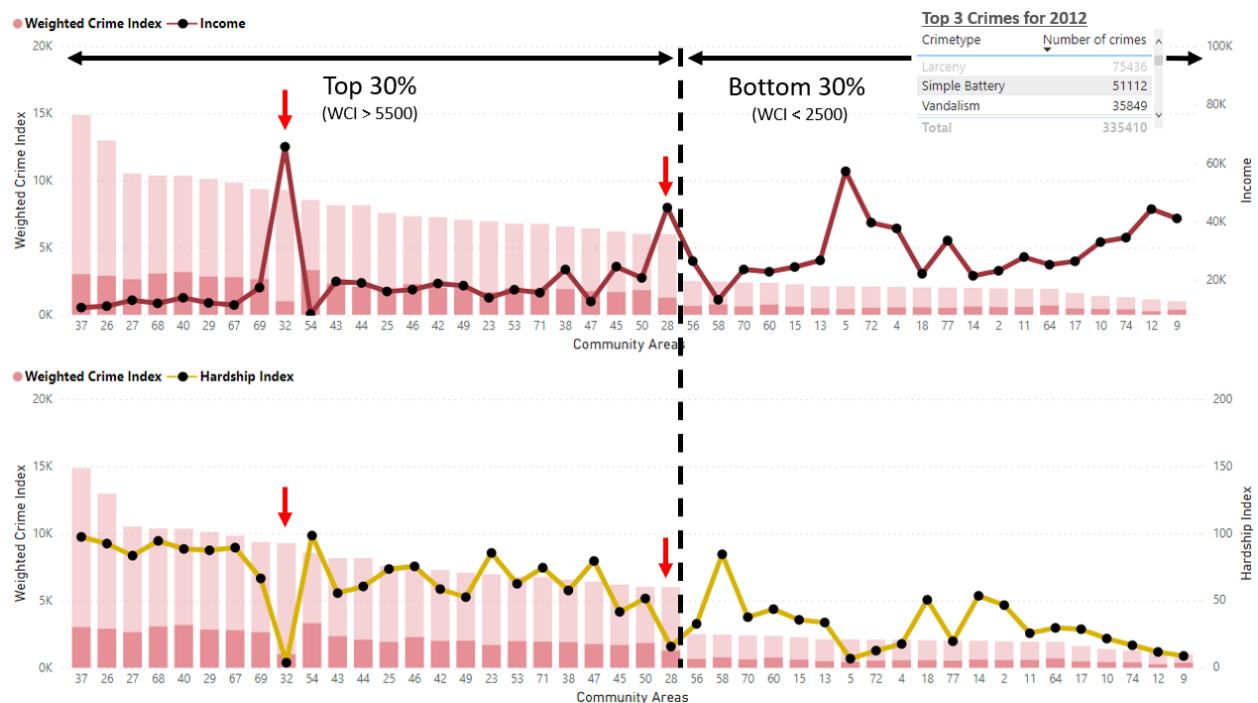


Figure 14 WCI, Income and Hardship Index per community area for batteries and vandalism

In Figure 12 we tried to add Income and Hardship Index to the equation by comparing 30% of the areas with the highest WCIs with 30% of the areas with the smallest WCIs, as a way to see if living in a rich neighborhood guarantees safety.

Initially the results were not as expected, as some of the top areas in criminality had high incomes and low hardship indexes. However, after inspecting the areas for the top 3 crimes (first for larceny in Figure 13 and then for battery and vandalism in Figure 14) it was found that the aforementioned areas, with high incomes and low hardship indexes, were scoring high in criminality only because of the high number of larcenies which took place there. Considering the above, there is an upward trend in the WCI Index as income decreases and the hardship index increases.

In the next page, (Figure 15), by inspecting the crime locations on the map, two clusters closely distributed to the west side of the city can be spotted. Our hypothesis, after checking [Google Maps Chicago Police Departments](#), is that the areas with the lowest WCIs are the ones with a police department in them, although we could not find an accurate dataset to incorporate in our warehouse. Of course, this hypothesis contradicts our earlier conclusion that CPD does not contribute adequately in fighting crime in the city.

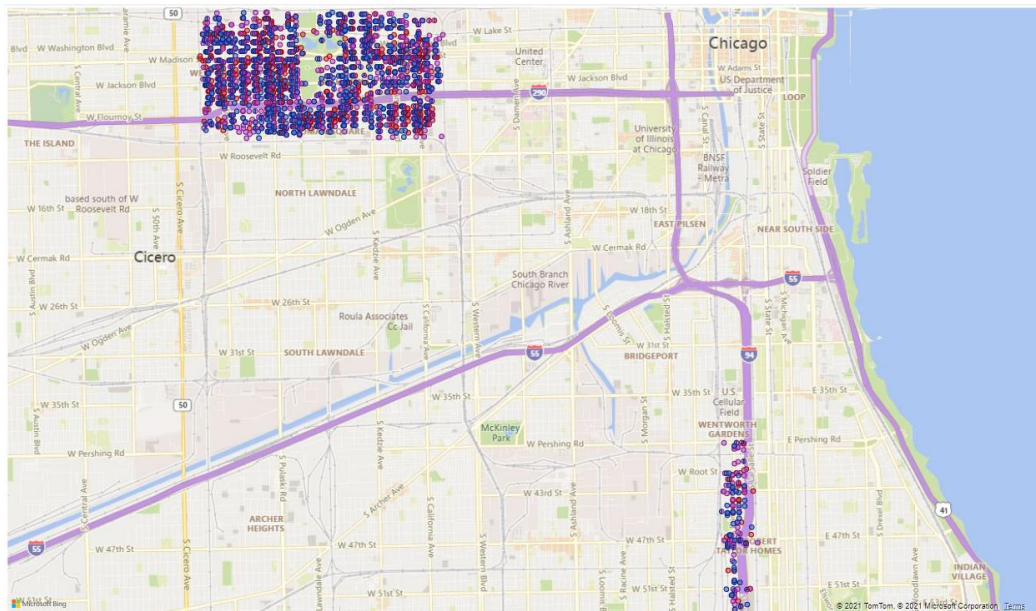


Figure 15 Top 3 Community Areas on the map

Upon further inspection, we confirmed that crime locations are actually building blocks and not exact addresses, which produces this grid-like appearance on the map pins' distribution. We also noticed that all three regions were traversed by a highway.

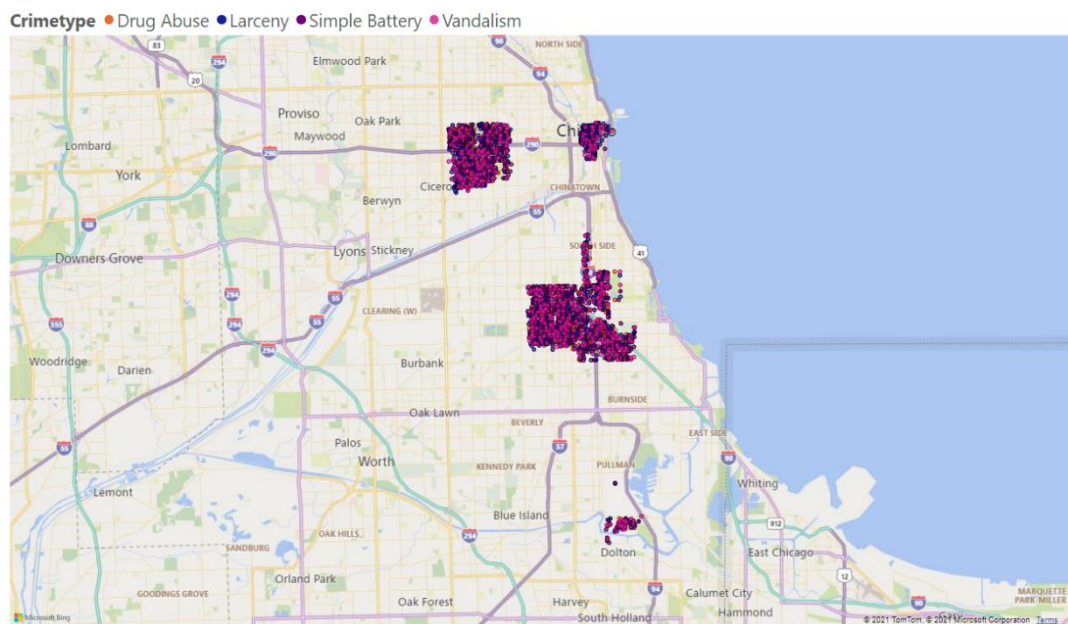


Figure 9 Top 10 Community Areas on the map

After expanding our search to the top 10 areas, as seen in *Figure 16*, we quickly refuted any correlation between highways and criminality rate, although the clusters we found in *Figure 11* were quite similar. We did, however, note that downtown Chicago (Loop) and the small Dolton suburban area of Riverdale had an unusually high WCI, even though they seemed completely out of place. For Loop, in addition to our previous findings about the high number of larcenies, we found that it's the 2nd largest commercial business district in North America (Wikipedia, 2021) and the busiest part of downtown Chicago, which would explain the high WCI out of sheer overcrowding. For Riverdale we could not reach any relevant conclusion for this discrepancy outside of demographic indices.

Com Areas	Com Areas Nms	Population	Weighted Crime Index
37	Fuller Park	2876	14,871.35
26	West Garfield Park	18001	12,962.89
27	East Garfield Park	20567	10,510.04
68	Englewood	30654	10,370.75
40	Washington Park	11717	10,350.35
29	North Lawndale	35912	10,114.45
67	West Englewood	35505	9,832.14
69	Greater Grand Crossing	32602	9,360.32
32	Loop	29283	9,278.59
54	Riverdale	6482	8,547.52
Total		223599	10,179.16

Finally, in *Figure 17* we attempted to see if the 5 crimes under study were diminishing over time. Evidently, there was a slight decrease in all of them.

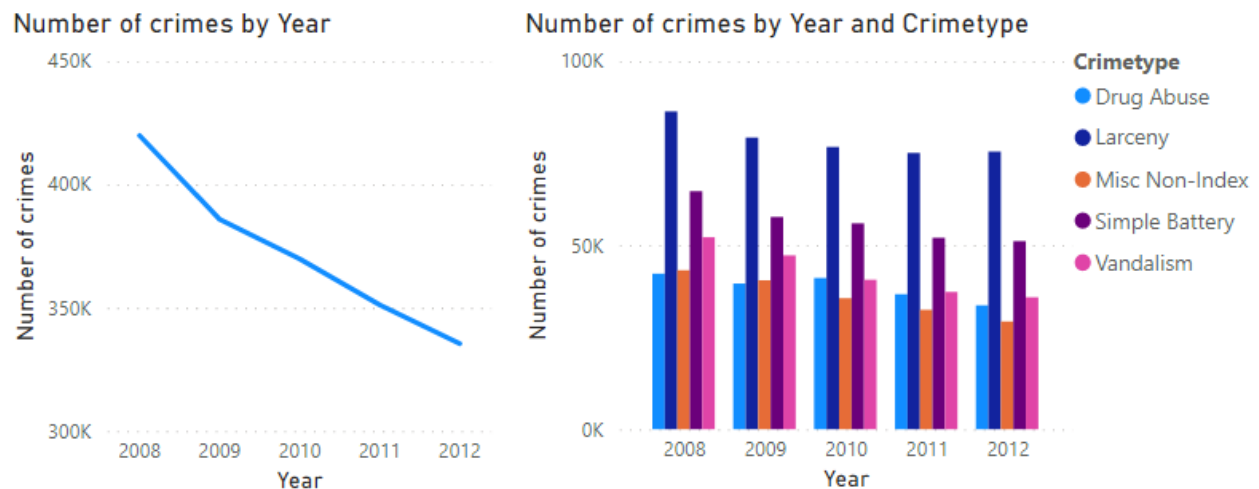


Figure 17 Top 5 crimes over time

6. Conclusions

Our study revolved around finding useful demographic factors and their connections to criminality rate when calculating prices and benefits for future insurance packages. The city of Chicago proved to be invaluable, as their data portal had a massive amount of data, which, after proper cleaning and transforming, we were able to use to test our hypotheses. Unfortunately, though, we were unable to focus in

more up-to-date data, due to most of the data portal's accurate datasets being over a decade old. Still, some very useful information was acquired by processing it.

The process included creating our own data warehouse, using various datasets and xls tables from the data portal, R scripts for data cleaning, SSMS's integration services for the database, SSDT's analysis services for the cube and Power BI for our visualizations and reports.

Our conclusions regarding the data, take into account our assumed role of an insurance company and are as follows:

- Criminality rate overall in Chicago city during the years 2008-2012 was on the decline.
- Chicago PD's efforts have not contributed adequately in overall crime reduction.
- Simply counting crime instances is not an adequate measure for the 77 community areas. For that, a Weighted Crime Index (per capita) is much more efficient.
- There are geographical clusters of areas where criminality rate is much higher and others where criminality rate is low. These clusters share common demographic indicators besides simply being next to one another.
- For the few areas that fall outside of the above categories, those with unusually high criminality rates were due to their being industrial or business zones with very high traffic and crowding. Those with unusually low criminality rates did not share common criteria and features and their demographics were normal. Our assumption is that these areas host Police Departments or Headquarters, which we could not confirm due to lack of data on CPD's locations.
- The top five crime categories committed in Chicago are larceny, simple battery, drug abuse, vandalism and miscellaneous crimes.
- The amount of larcenies in an area increases with the area's income, whereas batteries and drug abuse increase with the area's hardship index. Vandalism is relatively the same among all areas but sees a slight increase with an area's hardship index. Miscellaneous crimes are too vague a category to be correlated with demographic indices.
- The top 5 crimes show a tendency of decrease over time by the end of 2012.
- Crimes that induce loss of life are few enough to be overlooked as a potential risk.

After taking into consideration these conclusions, we decided that factoring in someone's home area is important when calculating a customized insurance policy, especially when it involves property or car insurance, theft protection, safety & health insurance, and by extension life insurance. Also, by finding a client's address, one may be able to approximately calculate the client's household socioeconomic situation and his tendencies or susceptibilities towards crime.

7. Sources & Tools

General Info:

Chicago. (2021, 12 10). Chicago City. Retrieved from Chicago.gov:

<https://www.chicago.gov/city/en.html>

Wikipedia. (2021, 12 10). *Chicago*. Retrieved from Wikipedia, the free encyclopedia:

<https://en.wikipedia.org/wiki/Chicago>

Wikipedia. (2021, 12 10). *Chicago, Loop*. Retrieved from Wikipedia, the free encyclopedia:

https://en.wikipedia.org/wiki/Chicago_Loop

Datasets:

Chicago Data Portal. (2021, 11 29) *Chicago Data Portal main*. Retrieved from CDP:

<https://data.cityofchicago.org/>

Crimes – 2001 to Present (2021, 11 29) *Public Safety*, Retrieved from CDP:

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2>

Census Data – Selected socioeconomic Indicators in Chicago, 2008-2012 (2021, 11 30) *Health & Human Services*, Retrieved from CDP:

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2>

Community Area 2000 and 2010 Census Population Comparisons (2021, 12 4) *Planning and Development*, Retrieved from Chicago.gov:

https://www.chicago.gov/city/en/depts/dcd/supp_info/community_area_2000and2010censuspopulationcomparisons.html

Tools:

RStudio

Excel

Microsoft SQL Server Management Studio 18

Visual Studio 2017 (SSDT) Standalone edition

Power BI Desktop