

# Práctica Final: Rodrigo López de Toledo

## Abril 2020

### Análisis del crimen en Chicago

#### 1. Introducción:

Chicago es una de las ciudades de Estados Unidos, y por extensión del mundo, con mayor índice de criminalidad. Los crímenes varían desde los mas sencillos a los mas inimaginables, y además, se ramifican desde la ciudad por todo el Midwest estadounidense debido a la localización céntrica de la ciudad.

El objetivo de este proyecto es hacer un estudio detallado de los crímenes mas comunes y mostrar en que zonas de la ciudad se deben destinar mas recursos para unos tipos de crímenes, bajo la suposición de que, si se pueden separar los crímenes por sus localizaciones mayoritarias, las campañas para reducirlos serán mas efectivas si se sabe el objetivo mas que si se hacen campañas generales.

El dataset disponible en el link (<https://www.kaggle.com/chicago/chicago-crime>) contiene datos de todo el siglo XXI. Como es una cantidad demasiado grande de datos, que mi ordenador no puede procesar, se va ha utilizar los años mas recientes. El análisis general será sobre los crímenes del año 2019.

Para la localización exacta de cada distrito he usado el siguiente dataset oficial del gobierno de Illinois: <https://data.cityofchicago.org/widgets/z8bn-74gv>

A pesar de hacer el análisis solo sobre el año 2019, hay diversos métodos que necesitan de un set de training y otro set de test para hacer las predicciones, como pueden ser arboles de decisión o clasificadores como KNN. En ese caso se utilizarán los datos de 2016 a 2018 como training y los de 2019 como test, para ver la precisión de los diversos clasificadores.

Para la exploración inicial de datos y métodos propios de data mining como puede ser encontrar las reglas de asociación, se usará simplemente el set de 2019.

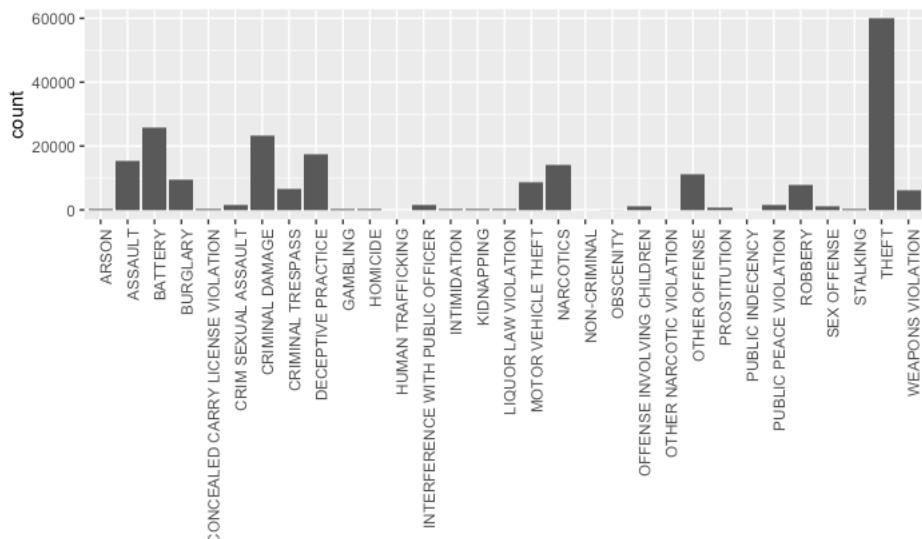
#### 2. Exploración de los datos:

Como se puede comprobar en el link proporcionado de la fuente del dataset, este tiene 22 columnas. Algunas columnas que se incluyen no son necesarias, o son redundantes, como pueden ser IUCR o fbi\_code. Por ello lo primero que vamos a hacer es eliminar esas columnas, quedándonos el dataset como se muestra en la siguiente imagen.

case_number	block	primary_type	description	location_description	arrest	district	ward	latitude	longitude	month	day
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	JC294267	002XX W GARFIELD BLVD	ASSAULT	PRO EMP HANDS NO/MIN INJURY	CTA STATION	True	9	41.79453	-87.63144	6	6
2	JC302504	095XX S PROSPECT AVE	OTHER OFFENSE	HARASSMENT BY ELECTRONIC MEANS	SCHOOL PUBLIC BUILDING	False	22	41.72055	-87.66536	6	11
3	JC341026	111XX S TROY ST	NARCOTICS	POSS COCAINE	RESIDENCE	True	22	41.69062	-87.69957	7	9
4	JC357542	059XX S MENARD AVE	NARCOTICS	MANU/DELCANNABIS OVER 10 GMS	RESIDENCE	True	8	41.78382	-87.76684	7	20
5	JC413526	063XX S LONG AVE	DECEPTIVE PRACTICE	BOGUS CHECK	OTHER	False	8	41.77692	-87.75682	7	27
6	JC437390	012XX E 80TH ST	NARCOTICS	POSS METHAMPHETAMINES	STREET	True	4	41.74956	-87.59497	9	17

Como se ha comentado antes, el objetivo es hacer un análisis de los crímenes por los distritos y por sus arrestos, para que las autoridades puedan hacer campañas específicas por distrito frente a un crimen u otro.

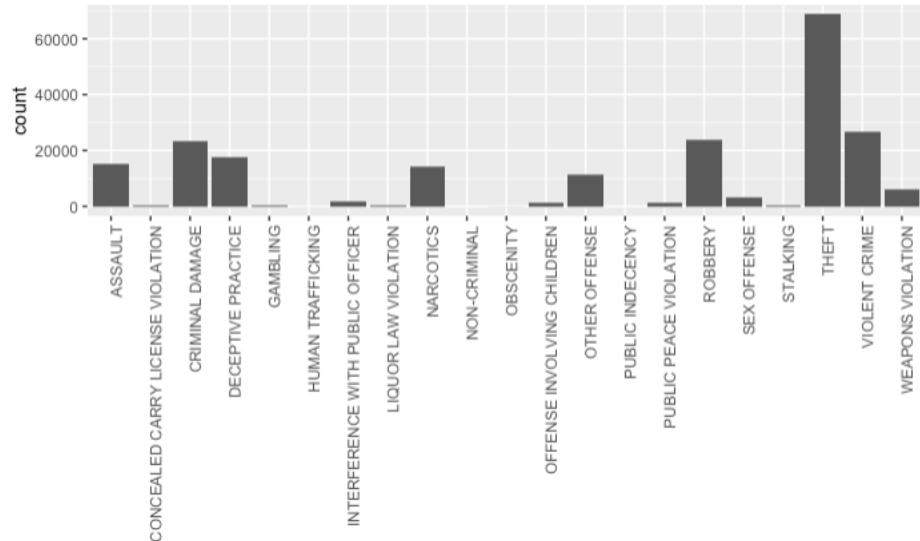
Para ello, vamos a ver primero como están distribuidos los crímenes durante el ultimo año.



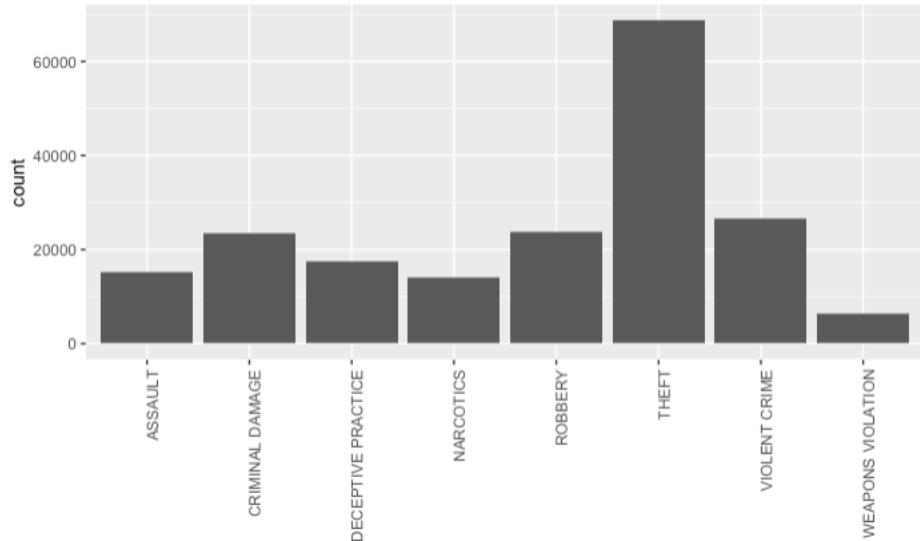
Como es lógico viendo esta distribución, me voy a centrar en los crímenes mas comunes, que es donde se necesita una intervención mas urgente. Antes de elegir los crímenes, creo necesario agrupar diversos crímenes muy similares para que la granularidad de este atributo no sea tan compleja. El motivo de estas agrupaciones es el siguiente, el objetivo de este proyecto no es clasificar hasta el mas mínimo detalle los crímenes, sino proporcionar una imagen general de donde se necesita mas intervención por un crimen similar. En la siguiente tabla se ven los cambios:

TRANSFORMACIONES	
CRIMINAL TRESPASS	ROBBERY
BURGLARY	ROBBERY
MOTOR VEHICLE THEFT	THEFT
HOMICIDE	VIOLENT CRIME
KIDNAPPING	VIOLENT CRIME
BATTERY	VIOLENT CRIME
INTIMIDATION	VIOLENT CRIME
ARSON	VIOLENT CRIME
PROSTITUTION	SEX OFFENSE
CRIM SEXUAL ASSAULT	SEX OFFENSE
OTHER NARCOTIC VIOLATION	NARCOTICS

Después de estas transformaciones los crímenes quedan así:

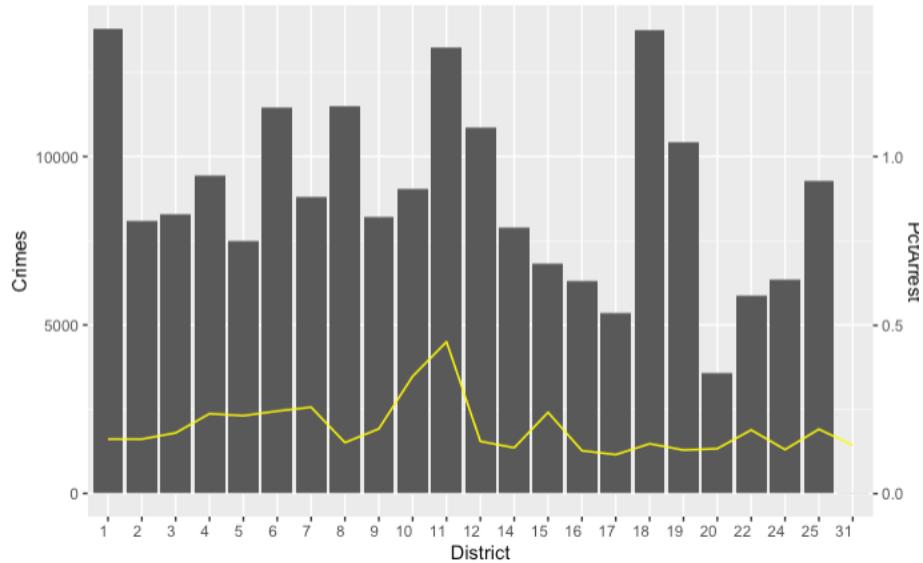


A la luz de estos resultados los crímenes en los que nos centraremos son los siguientes, mostrados en la figura con sus correspondientes casos.



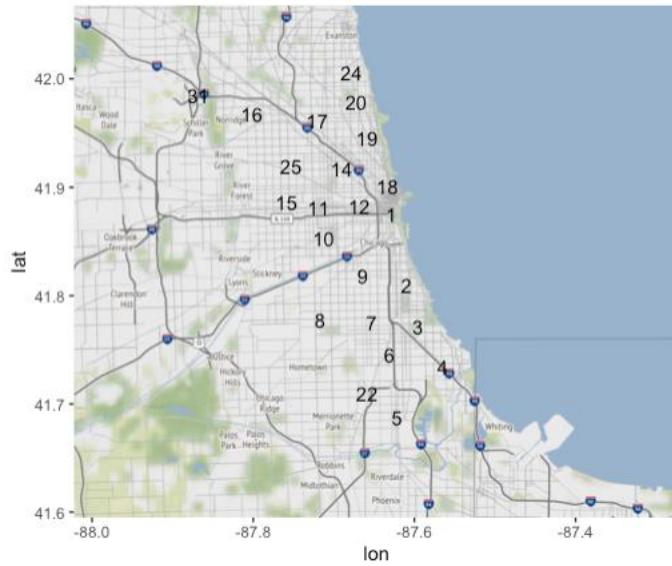
Estos son los crímenes que he decidido que son los mas importantes.

A continuación, voy a explorar los datos que tenemos relacionados con los distritos, a ver si puedo empezar a plantear alguna hipótesis inicial.



Como se aprecia, la distribución de crímenes por distrito esta razonablemente equilibrada, habiendo a pesar de ello alguna característica digna de mencionar, como puede ser que los distritos 1, 11 y 18 parecen ser que son los que mas problemas de crímenes tienen, y siendo, a priori, el 17 y el 20 los mas seguros. Vamos a ver el mapa a ver si hay alguna razón para ello.

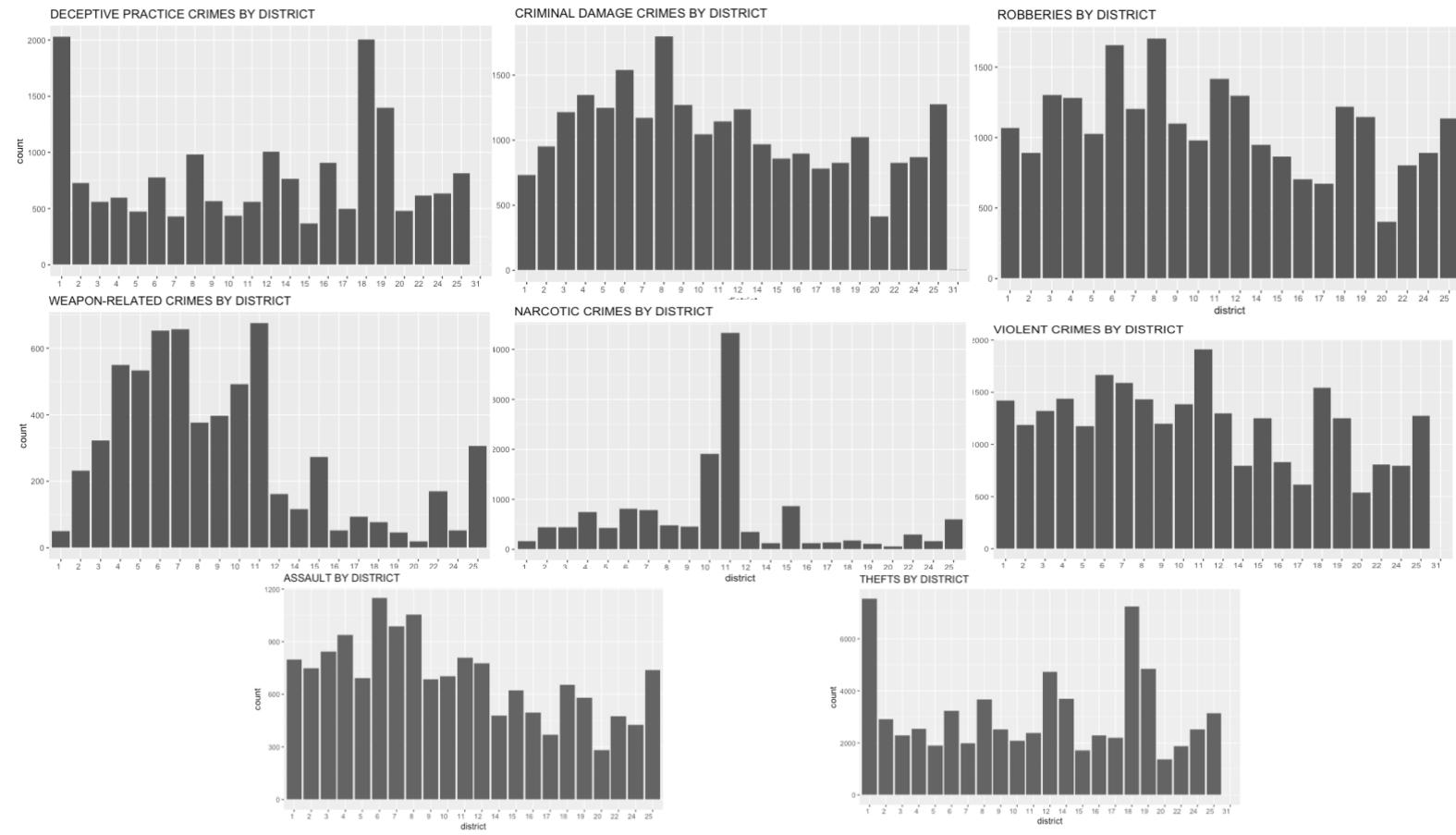
La línea amarilla muestra el porcentaje de arrestos por distrito, indicando que es bastante bajo para todos los distritos, pero indicando claramente como en el distrito 10 y 11, el porcentaje es mayor, por lo que habrá que estar atento a lo que ocurre con esos distritos a lo largo de este estudio.



Como se ve en el mapa, los distritos 1 y 18 corresponden con el centro de la ciudad, donde hay tanto mas densidad de gente como donde mas gente pasa al día, pudiendo ser ello una posible causa para el alto número de crímenes, además este motivo no justifica la criminalidad en el distrito 11. Los distritos 17 y 20, se sitúan en los suburbios del norte,

correspondiente a urbanizaciones para gente con mayor poder adquisitivo y, por lo tanto, mas seguros.

A continuación, voy a mostrar por cada tipo de crimen donde ocurren.

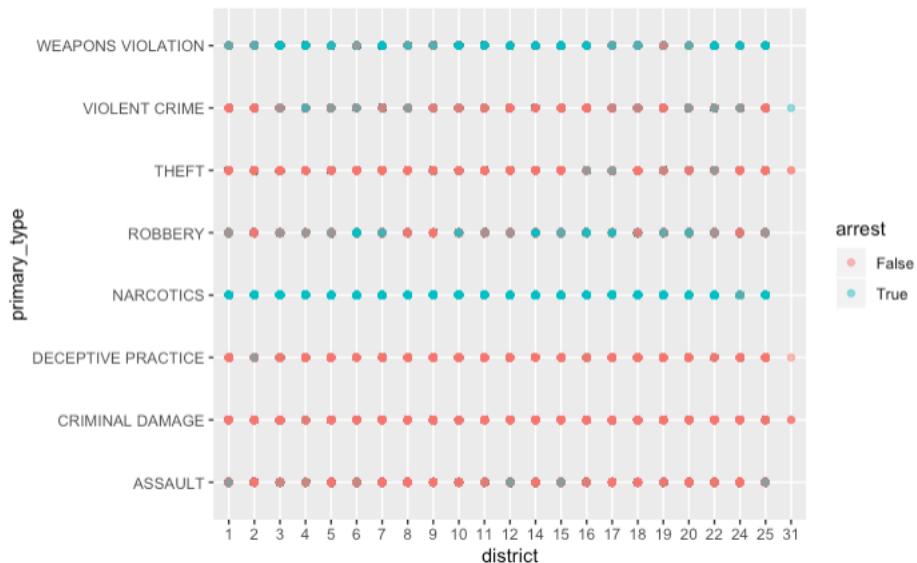


Viendo las gráficas anteriores, se pueden empezar a lanzar varias hipótesis. Tanto DECEPTIVE CRIMES como THEFTS ocurren en un número bastante mayor al resto de distritos 1 y 18 (Recordemos, el centro de la ciudad, donde se concentra mas gente y la renta per cápita es mayor)

WEAPONS VIOLATIONS ocurren con mucha mas frecuencia en distritos situados al sur de la ciudad (Cabe aclarar, que, en Chicago, es muy conocido que la zona sur y oeste de la ciudad es significativamente mas peligrosa, y pobre que la zona este y norte). Lo mismo ocurre con ROBBERIES, ASSAULT y CRIMINAL DAMAGE, pero en estos casos esta tendencia es bastante menos definida.

En cuanto a VIOLENT CRIMES esta bastante mas repartido, excepto un alto numero en el distrito 11, que si nos fijamos en NARCOTICS es, con mucha diferencia, el distrito donde mas crímenes de drogas se producen, siendo esta hipótesis la que mas fuerte parece.

Por último, vamos a ver la correlación entre estas tres variables sobre las que estamos poniendo el foco, primary\_type, arrest y district.



En la imagen, con un gradiente de rojo a azul turquesa se muestra el porcentaje de arrestos por crimen y distrito. Se pueden ver tendencias claras como, por ejemplo, la claridad con la que se ve que los arrestos con crímenes de narcóticos son muy comunes, así como en crímenes de armas, aunque sean menos comunes. Hay crímenes en los que independientemente del distrito, el arresto es casi siempre falso, como puede ser ASSAULT, CRIMINAL DAMAGE o DECEPTIVE PRACTICE, y otros en los que depende del distrito donde se produzca, como ROBBERY o en menor medida, VIOLENT CRIME.

Estas hipótesis se tratarán de demostrar o al menos de poder considerarlas válidas con los métodos siguientes que vamos a aplicar.

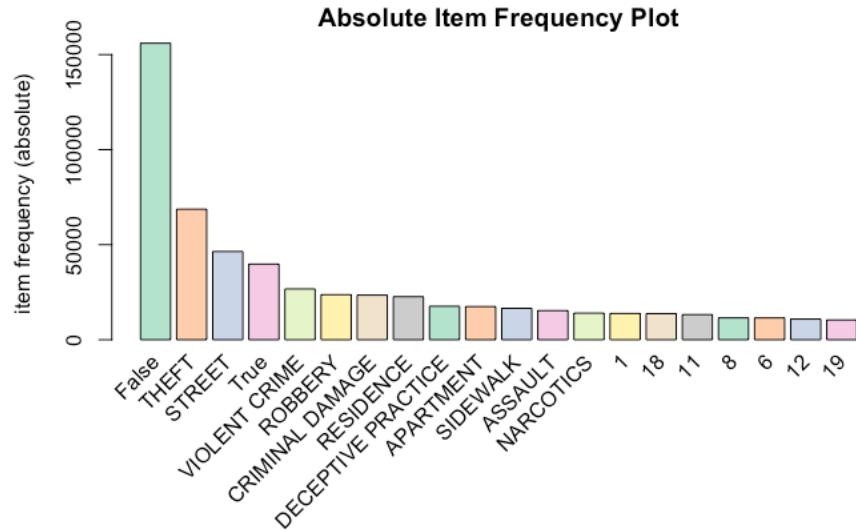
### 3. Reglas de Asociación

En este apartado vamos a aplicar la técnica de reglas de asociación para tratar de probar de una manera mas clara e inequívoca algunas de las hipótesis contadas arriba. Algunas de esas hipótesis podrán ser probadas con esta técnica, pero otras no, por lo que nos será muy útil para ver cuales son mas posibles según avance el estudio. Esta técnica de reglas de asociación utiliza un algoritmo de data mining muy conocido, llamado a priori, pero realmente, esta técnica no es mas que un método para encontrar relaciones mas o menos frecuentes en el dataset, por lo que es casi una continuación de la exploración de datos.

Para comenzar con este análisis, al no ser una técnica de clasificación o predicción, se puede prescindir de muchas variables, ya que las relaciones que mas nos interesan son aquellas entre los distritos, los crímenes y los arrestos. Por ello, voy a quedarme solo con las variables *primary\_type*, *arrest*, *district* y *location\_description*.

Sobre todos los pasos y gráficos que se comenten, se hará un análisis general de todo el dataset y un análisis de cada crimen en concreto.

Haciendo un análisis de la frecuencia con la que aparecen los ítems obtenemos el gráfico de barras siguiente. Por la frecuencia de cada ítem, parece que encontraremos relaciones bastante frecuentes con el resultado de no arresto, así como algún tipo de crimen como THEFT, en los distritos más afectados que son el 1, 18, 11, 8, 6, así como un gran número de crímenes en aceras (SIDEWALKS), calles (STREET) o residencias (RESIDENCE).



Aplicando apriori una primera vez sobre el dataset general obtenemos las siguientes reglas que cumplen el requisito de más de un 70% de confianza. La primera lista muestra las reglas ordenadas por numero total de ocasiones que se produce esa asociación y la segunda en orden de confianza.

	<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{}	=>	{False}	0.796711548	0.7967115	1.000000	155977
[2]	{NARCOTICS}	=>	{True}	0.071449003	0.9994284	4.916430	13988
[3]	{DEPARTMENT STORE}	=>	{THEFT}	0.018521167	0.7557316	2.154410	3626
[4]	{11,SIDEWALK}	=>	{True}	0.012197614	0.7646494	3.761495	2388
[5]	{STREET,WEAPONS VIOLATION}	=>	{True}	0.011155606	0.7255814	3.569311	2184
[6]	{GROCERY FOOD STORE,True}	=>	{THEFT}	0.005971110	0.7814171	2.227633	1169
[7]	{SMALL RETAIL STORE,True}	=>	{THEFT}	0.005490969	0.7072368	2.016163	1075
[8]	{1,SMALL RETAIL STORE}	=>	{THEFT}	0.003713428	0.7127451	2.031866	727
[9]	{18,SMALL RETAIL STORE}	=>	{THEFT}	0.003703212	0.7149901	2.038266	725
[10]	{BANK}	=>	{DECEPTIVE PRACTICE}	0.003432494	0.7044025	7.825291	672

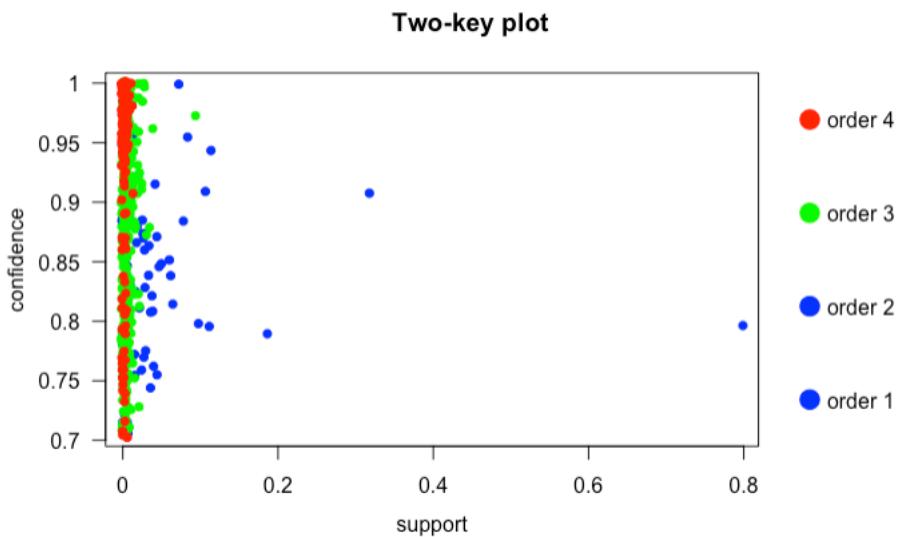
	<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[11]	{11,WEAPONS VIOLATION}	=>	{True}	0.002778686	0.8059259	3.964545	544
[12]	{7,WEAPONS VIOLATION}	=>	{True}	0.002712283	0.8082192	3.975826	531
[13]	{6,WEAPONS VIOLATION}	=>	{True}	0.002410919	0.7239264	3.561169	472
[14]	{4,WEAPONS VIOLATION}	=>	{True}	0.001992073	0.7090909	3.488190	390
[15]	{ATHLETIC CLUB}	=>	{THEFT}	0.001895023	0.7745303	2.208001	371
[16]	{POLICE FACILITY/VEH PARKING LOT}	=>	{True}	0.001874591	0.7399194	3.639842	367
[17]	{10,WEAPONS VIOLATION}	=>	{True}	0.001782650	0.7093496	3.489462	349
[18]	{19,RESIDENCE PORCH/HALLWAY}	=>	{THEFT}	0.001389343	0.7513812	2.142008	272
[19]	{ATM (AUTOMATIC TELLER MACHINE)}	=>	{DECEPTIVE PRACTICE}	0.001363804	0.8530351	9.476469	267
[20]	{18,GROCERY FOOD STORE}	=>	{THEFT}	0.001328048	0.7084469	2.019613	260

<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {NARCOTICS}	=>	{True}	0.071449003	0.9994284	4.916430	13988
[2] {VEHICLE NON-COMMERCIAL,WEAPONS VIOLATION}	=>	{True}	0.001006252	0.9425837	4.636798	197
[3] {ATM (AUTOMATIC TELLER MACHINE)}	=>	{DECEPTIVE PRACTICE}	0.001363804	0.8530351	9.476469	267
[4] {7,WEAPONS VIOLATION}	=>	{True}	0.002712283	0.8082192	3.975826	531
[5] {11,WEAPONS VIOLATION}	=>	{True}	0.002778686	0.8059259	3.964545	544
[6] {}	=>	{False}	0.796711548	0.7967115	1.000000	155977
[7] {CTA STATION,DECEPTIVE PRACTICE}	=>	{True}	0.001133949	0.7900356	3.886376	222
[8] {GROCERY FOOD STORE,TRUE}	=>	{THEFT}	0.005971110	0.7814171	2.227633	1169
[9] {ATHLETIC CLUB}	=>	{THEFT}	0.001895023	0.7745303	2.208001	371
[10] {11,SIDEWALK}	=>	{True}	0.012197614	0.7646494	3.761495	2388

Hay reglas muy interesantes que se pueden extraer de aquí, la primera es el gran porcentaje de arresto que se producen en crímenes de NARCOTICS, teniendo en cuenta que en un casi 80% de los casos totales no se producen arrestos, así como en el distrito 11 si el crimen es en la acera.

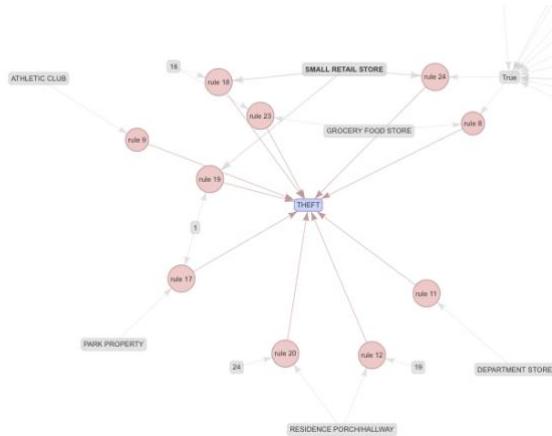
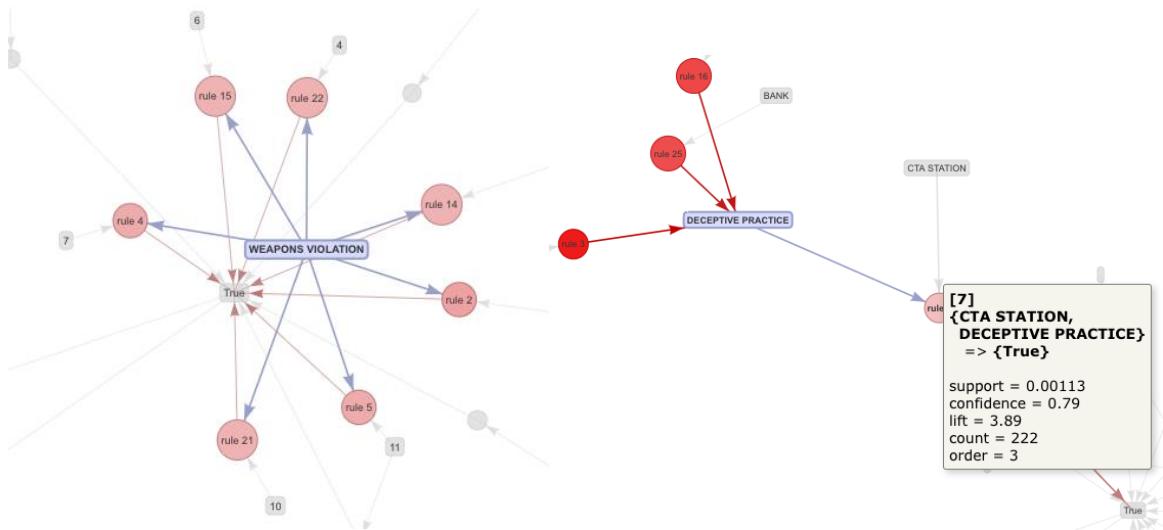
Otras reglas parecen relacionar mucho los distritos 1 y 18 con THEFT, así como los distritos 7,6,4 y 11 con WEAPONS VIOLATION, teniendo este, al parecer, un alto porcentaje de arrestos.

A continuación, se muestran los diferentes gráficos para analizar el dataset completo.

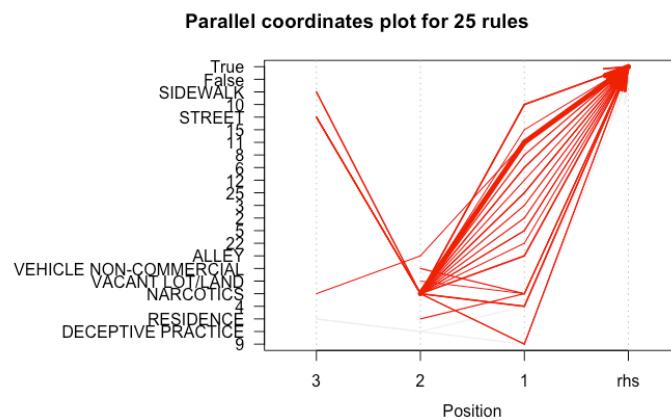


En este primer grafico se ve la distribución de las reglas en función del support y la confianza. Como se ve la mayoría tienen muy poco support, debido a que el dataset es muy grande. Por ello, he preferido mostrar las reglas tanto por número de casos totales, como por la confianza, que considero métricas mas apropiadas para este estudio.

Haciendo una representación de grafos se ve claramente como para WEAPONS VIOLATIONS, los distritos mas afectados son el 4,6,7, 10 y 11, además de tener un porcentaje muy alto de arrestos, similar a cuando se produce un robo (THEFT) en una tienda.



Si hacemos una grafica de coordenadas se ve como los distritos 10 y 11 son los mas problemáticos en cuanto a NARCOTICS, teniendo además un gran número de arrestos.



Ahora pasamos a buscar relaciones, pero centrándome solo en cada crimen en particular, a ver si puedo sacar mas conclusiones. En algunos crímenes será necesario limpiar aun mas los datos, para poder extraer algo de conocimiento.

- **Assault:**

De nuevo, muestro primero las reglas por el numero total de casos y luego por confianza.

lhs <fctr>	<fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {SIDEWALK}	=>	{ASSAULT}	0.008974542	0.1063946	1.360605	1757
[2] {6}	=>	{ASSAULT}	0.005863844	0.1001221	1.280391	1148
[3] {7}	=>	{ASSAULT}	0.005041476	0.1119682	1.431883	987
[4] {3}	=>	{ASSAULT}	0.004305941	0.1016520	1.299956	843
[5] {SCHOOL PUBLIC BUILDING}	=>	{ASSAULT}	0.002993217	0.2306179	2.949209	586
[6] {GAS STATION}	=>	{ASSAULT}	0.002257682	0.1294290	1.655176	442
[7] {RESIDENCE PORCH/HALLWAY}	=>	{ASSAULT}	0.002181064	0.1011369	1.293368	427
[8] {ALLEY,False}	=>	{ASSAULT}	0.001614090	0.1150346	1.471096	316
[9] {APARTMENT,True}	=>	{ASSAULT}	0.001384235	0.1344246	1.719061	271
[10] {RESIDENCE,True}	=>	{ASSAULT}	0.001210567	0.1139423	1.457128	237

lhs <fctr>	<fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {SCHOOL PUBLIC BUILDING}	=>	{ASSAULT}	0.002993217	0.2306179	2.949209	586
[2] {APARTMENT,True}	=>	{ASSAULT}	0.001384235	0.1344246	1.719061	271
[3] {GAS STATION}	=>	{ASSAULT}	0.002257682	0.1294290	1.655176	442
[4] {ALLEY,False}	=>	{ASSAULT}	0.001614090	0.1150346	1.471096	316
[5] {RESIDENCE,True}	=>	{ASSAULT}	0.001210567	0.1139423	1.457128	237
[6] {7}	=>	{ASSAULT}	0.005041476	0.1119682	1.431883	987
[7] {SIDEWALK}	=>	{ASSAULT}	0.008974542	0.1063946	1.360605	1757
[8] {CONVENIENCE STORE}	=>	{ASSAULT}	0.001093086	0.1024414	1.310050	214
[9] {3}	=>	{ASSAULT}	0.004305941	0.1016520	1.299956	843
[10] {RESIDENCE PORCH/HALLWAY}	=>	{ASSAULT}	0.002181064	0.1011369	1.293368	427

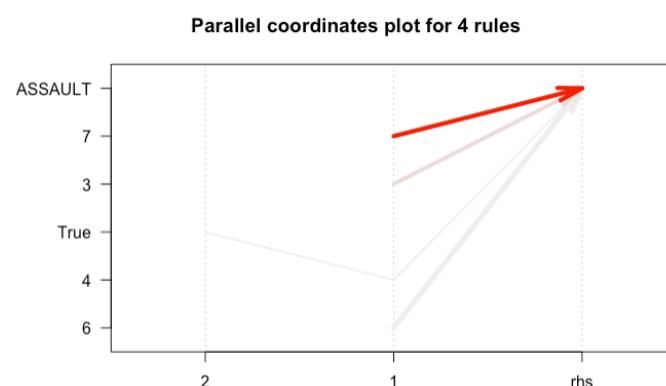
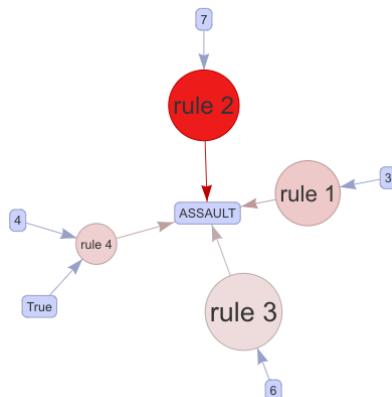
Mirando las tablas, no se ve ninguna regla que nos haga pensar que se puede extraer información 100% relevante y cierta sobre este crimen mas que parece que los distritos 6,7 y 3 parecen ser los mas afectados, pero la confianza es demasiado pequeña para asumirlo con certeza.

Sin embargo, si filtramos las localizaciones y nos quedamos con distritos y arrestos, parece que la conjectura de que los barrios mas afectados son 3,6 y 7 se hace mas fuerte. Además, hay una regla en particular muy interesante, indicando que en el barrio 4, la probabilidad de arresto es mas alta que la de no arresto, algo contrario a lo que se ve en las otras. Ciertamente es que el numero de eventos asi es muy bajo, pero es algo a lo que debería prestarse atención.

lhs <fctr>	<fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {6}	=>	{ASSAULT}	0.005863844	0.1001221	1.280391	1148
[2] {7}	=>	{ASSAULT}	0.005041476	0.1119682	1.431883	987
[3] {6,False}	=>	{ASSAULT}	0.004505149	0.1018241	1.302156	882
[4] {3}	=>	{ASSAULT}	0.004305941	0.1016520	1.299956	843
[5] {7,False}	=>	{ASSAULT}	0.004035224	0.1204636	1.540523	790
[6] {3,False}	=>	{ASSAULT}	0.003601054	0.1036460	1.325455	705
[7] {4,True}	=>	{ASSAULT}	0.001154381	0.1011186	1.293134	226

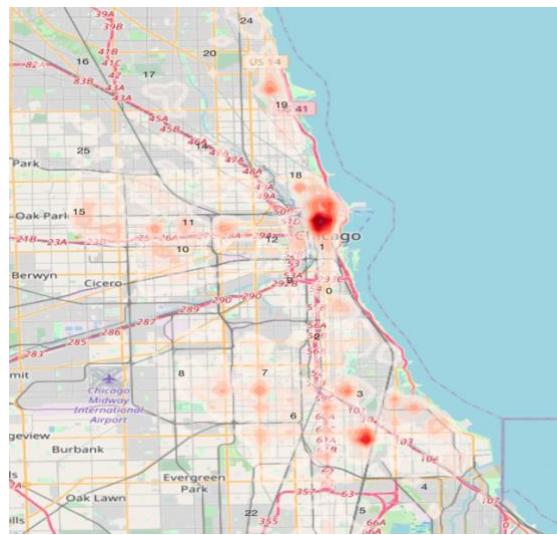
Aplicando nuevas representaciones a estas reglas, confirma lo que acabo de contar.

Indicando que el distrito mas afectado de esta pequeña selección sea el 7.



En el siguiente mapa, se ve una distribución de los asaltos en la ciudad. Antes de comentar lo que se ve, debo comentar que la gran densidad en el centro de la ciudad puede distorsionar el resultado, es decir, la gran densidad de casos en el centro se debe a que en el centro de la ciudad hay mucha mas gente y la zona es mas pequeña, pero eso no nos debe hacer pensar que en el resto de barrios no hay ese problema, de hecho si vemos la exploracion de datos, los distritos 1 y 18 no aparecen como los mas problematicos en este crimen.

Por lo tanto, viendo el siguiente mapa, si parece que hay una densidad mayor en la zona de los distritos 3,4 y 5, en el resto, quitando el centro de la ciudad, no parece que haya tanta incidencia.



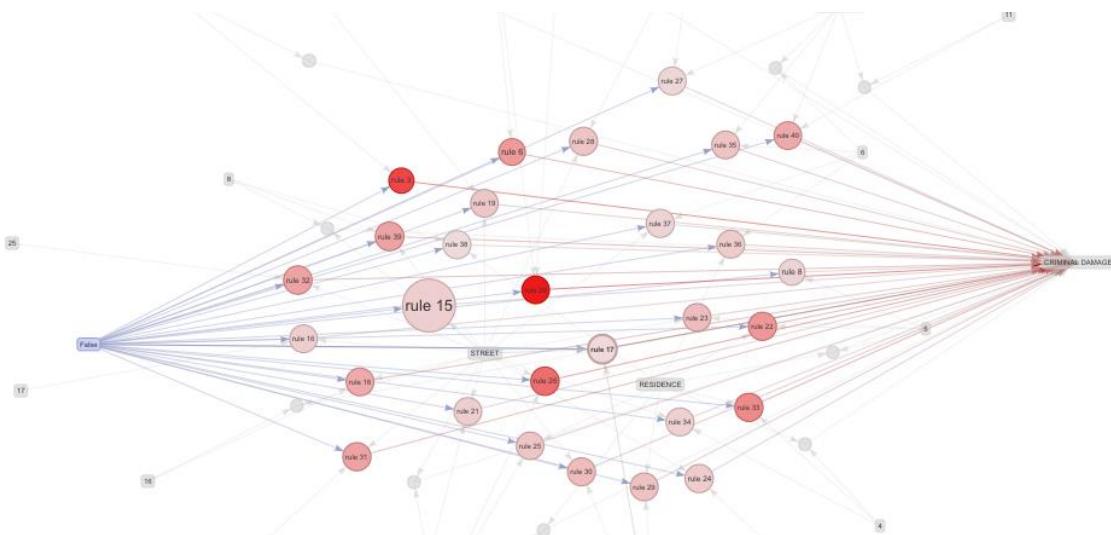
#### - Criminal Damage:

Para este crimen, de las tablas no se puede extraer ninguna conclusión. La discrepancia entre numero de casos, confianza y gran variedad entre los distritos implicados hace que no se pueda extraer ninguna conclusión mas que son barrios del sur - suroeste de la ciudad, además, para este crimen, su índice de arrestos es muy bajo.

	<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{False,STREET}	=>	{CRIMINAL DAMAGE}	0.039202967	0.2097625	1.751758	7675
[2]	{PARKING LOT/GARAGE(NON.RESID.)}	=>	{CRIMINAL DAMAGE}	0.007355345	0.2118893	1.769520	1440
[3]	{False,PARKING LOT/GARAGE(NON.RESID.)}	=>	{CRIMINAL DAMAGE}	0.007099951	0.2471111	2.063662	1390
[4]	{5,False}	=>	{CRIMINAL DAMAGE}	0.006042620	0.2052750	1.714282	1183
[5]	{RESIDENCE-GARAGE}	=>	{CRIMINAL DAMAGE}	0.004341697	0.2714788	2.267160	850
[6]	{False,RESIDENCE-GARAGE}	=>	{CRIMINAL DAMAGE}	0.004316157	0.2825142	2.359318	845
[7]	{8,STREET}	=>	{CRIMINAL DAMAGE}	0.003023864	0.2111270	1.763153	592
[8]	{8,False,STREET}	=>	{CRIMINAL DAMAGE}	0.002834873	0.2404679	2.008184	555
[9]	{9,STREET}	=>	{CRIMINAL DAMAGE}	0.002492645	0.2244710	1.874591	488
[10]	{25,False,STREET}	=>	{CRIMINAL DAMAGE}	0.002467105	0.2411383	2.013782	483

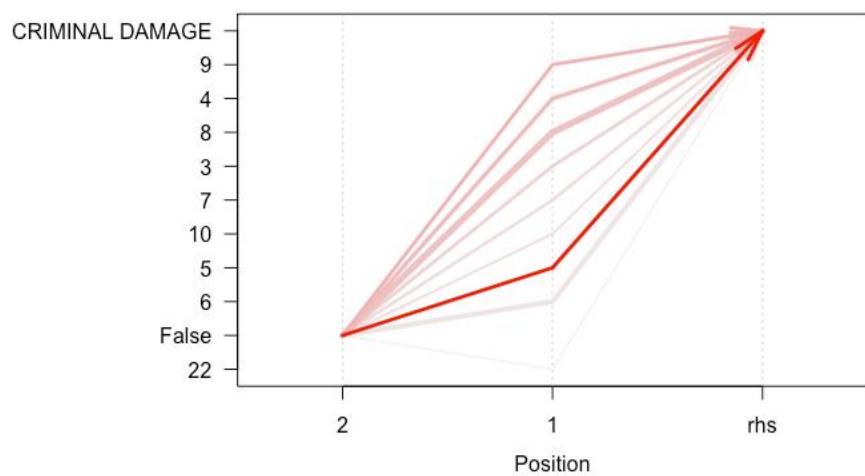
<b>lhs</b> <fctr>	<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {1, False, PARKING LOT/GARAGE(NON.RESID.)}	=> {CRIMINAL DAMAGE}	0.001128841	0.2990528	2.497434	221
[2] {1, PARKING LOT/GARAGE(NON.RESID.)}	=> {CRIMINAL DAMAGE}	0.001159488	0.2826899	2.360786	227
[3] {False, RESIDENCE-GARAGE}	=> {CRIMINAL DAMAGE}	0.004316157	0.2825142	2.359318	845
[4] {RESIDENCE-GARAGE}	=> {CRIMINAL DAMAGE}	0.004341697	0.2714788	2.267160	850
[5] {9, False, STREET}	=> {CRIMINAL DAMAGE}	0.002364948	0.2662450	2.223452	463
[6] {4, False, RESIDENCE}	=> {CRIMINAL DAMAGE}	0.002109554	0.2544670	2.125092	413
[7] {5, False, RESIDENCE}	=> {CRIMINAL DAMAGE}	0.002007396	0.2488917	2.078532	393
[8] {False, PARKING LOT/GARAGE(NON.RESID.)}	=> {CRIMINAL DAMAGE}	0.007099951	0.2471111	2.063662	1390
[9] {10, False, STREET}	=> {CRIMINAL DAMAGE}	0.002114662	0.2412587	2.014788	414
[10] {25, False, STREET}	=> {CRIMINAL DAMAGE}	0.002467105	0.2411383	2.013782	483

A pesar de ello, si vemos las gráficas resultantes, si parece que se puede extraer alguna conclusión.



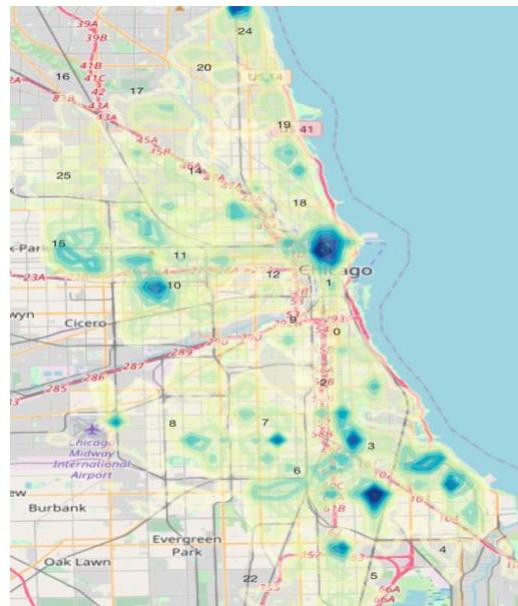
Como se puede ver en esta imagen, la cantidad de reglas inferidas de la relación FALSE con el crimen CRIMINAL DAMAGE es demasiado grande, por lo que podemos asumir que este crimen es uno de los mas bajos en cuanto a índice de arrestos.

Parallel coordinates plot for 10 rules



Limpiando la descripción de localizaciones, la gráfica anterior, confirma el reparto geográfico de los crímenes, así como la baja tasa de arrestos, añadiendo que puede que el 5 sea el mas afectado de todos ellos.

Viendo el mapa siguiente, se confirma el reparto mayoritario de los crímenes en los distritos indicados, sobretodo en los del sur de la ciudad, habiendo mas densidad en la zona de los distritos 3,4,5 y 6. A esto habría que añadir el 8, que en la exploración de datos aparece como el mas afectado, al ser uno de los mas grandes, la densidad no es mucha, pero el número total si. Este es otro caso que la densidad en el centro de la ciudad engaña si solo vemos este mapa, sin el análisis anterior.



- **Deceptive Practice:**

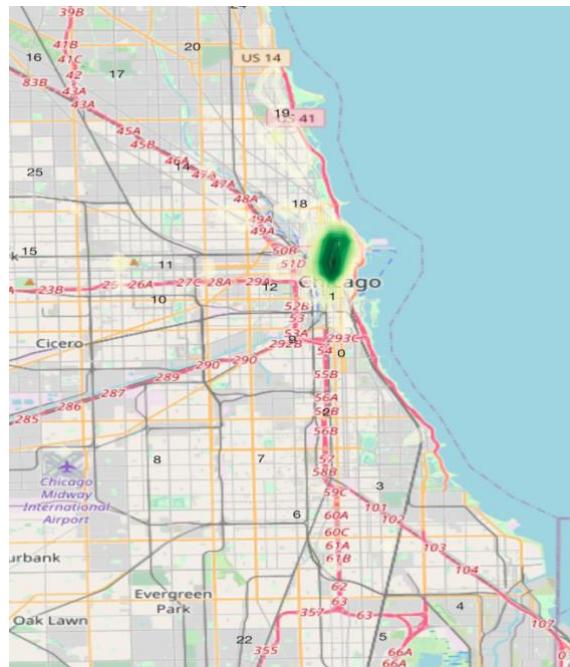
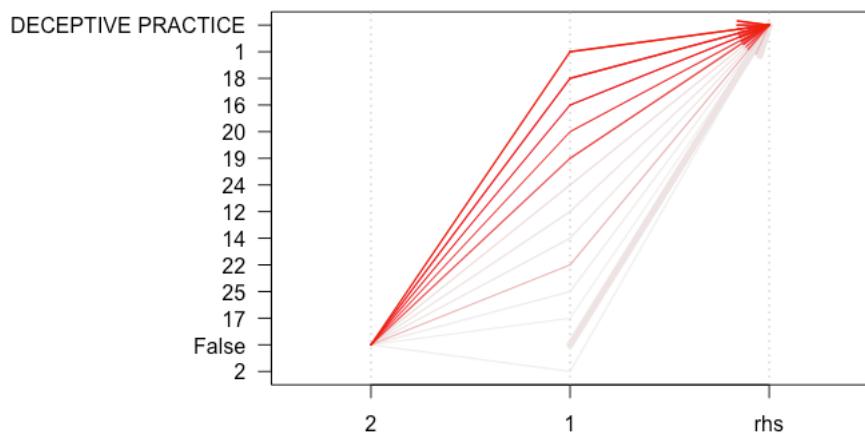
Mirando las siguientes tablas si parece claro los distritos mas afectados por este crimen, siendo estos el distrito 1,18 y 19. Además de los 17623 casos de DECEPTIVE PRACTICE totales, 16810 fueron saldados sin arresto.

Ihs <fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {False}	=> {DECEPTIVE PRACTICE}	0.085863436	0.1077723	1.197255	16810
[2] {RESIDENCE}	=> {DECEPTIVE PRACTICE}	0.026892980	0.2324709	2.582546	5265
[3] {OTHER}	=> {DECEPTIVE PRACTICE}	0.012366174	0.2807933	3.119366	2421
[4] {1}	=> {DECEPTIVE PRACTICE}	0.010358777	0.1469246	1.632203	2028
[5] {18}	=> {DECEPTIVE PRACTICE}	0.010236188	0.1458621	1.620400	2004
[6] {19}	=> {DECEPTIVE PRACTICE}	0.007135706	0.1342366	1.491250	1397
[7] {16}	=> {DECEPTIVE PRACTICE}	0.004627738	0.1438324	1.597851	906
[8] {RESTAURANT}	=> {DECEPTIVE PRACTICE}	0.004214000	0.1261661	1.401594	825
[9] {BANK}	=> {DECEPTIVE PRACTICE}	0.003432494	0.7044025	7.825291	672
[10] {DEPARTMENT STORE}	=> {DECEPTIVE PRACTICE}	0.003166885	0.1292205	1.435526	620

<b>lhs</b> <fctr>		<b>rhs</b> <fctr>		<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{ATM (AUTOMATIC TELLER MACHINE)}	=>	{DECEPTIVE PRACTICE}	0.001363804	0.8530351	9.476469	267
[2]	{CURRENCY EXCHANGE}	=>	{DECEPTIVE PRACTICE}	0.001225891	0.7228916	8.030688	240
[3]	{BANK}	=>	{DECEPTIVE PRACTICE}	0.003432494	0.7044025	7.825291	672
[4]	{OTHER}	=>	{DECEPTIVE PRACTICE}	0.012366174	0.2807933	3.119366	2421
[5]	{CTA STATION}	=>	{DECEPTIVE PRACTICE}	0.001435314	0.2666034	2.961729	281
[6]	{RESIDENCE}	=>	{DECEPTIVE PRACTICE}	0.026892980	0.2324709	2.582546	5265
[7]	{COMMERCIAL / BUSINESS OFFICE}	=>	{DECEPTIVE PRACTICE}	0.001532364	0.2024291	2.248809	300
[8]	{1}	=>	{DECEPTIVE PRACTICE}	0.010358777	0.1469246	1.632203	2028
[9]	{18}	=>	{DECEPTIVE PRACTICE}	0.010236188	0.1458621	1.620400	2004
[10]	{16}	=>	{DECEPTIVE PRACTICE}	0.004627738	0.1438324	1.597851	906

Esa tendencia se muestra en la siguiente imagen, pero, aparecen también muchos más barrios por lo que igual no es tan segura.

Parallel coordinates plot for 19 rules



En este caso en particular el análisis por reglas es bastante engañoso. El hecho de la baja tasa de arrestos nos engaña en los barrios por los que se reparte, parece que es un crimen

bastante extendido por la ciudad, pero eso no es cierto, viendo el mapa y la exploración de datos, en este caso, si que el centro de la ciudad es la zona mas afectada por este crimen.

- **Narcotics:**

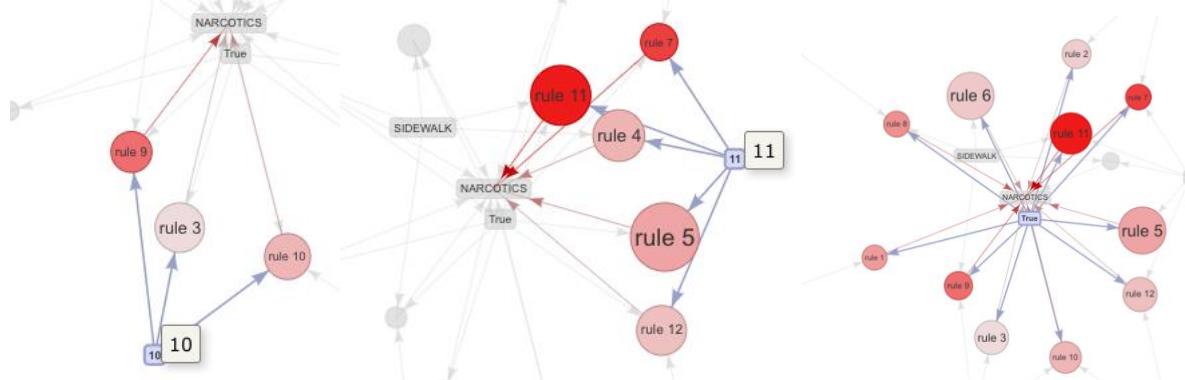
lhs <fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {True}	=> {NARCOTICS}	0.071449003	0.3514749	4.916430	13988
[2] {STREET}	=> {NARCOTICS}	0.025570039	0.1078600	1.508746	5006
[3] {STREET,True}	=> {NARCOTICS}	0.025564931	0.5095185	7.127143	5005
[4] {SIDEWALK}	=> {NARCOTICS}	0.022203947	0.2632312	3.682077	4347
[5] {SIDEWALK,True}	=> {NARCOTICS}	0.022193732	0.6482172	9.067260	4345
[6] {11}	=> {NARCOTICS}	0.022106898	0.3273580	4.579082	4328
[7] {11,True}	=> {NARCOTICS}	0.022106898	0.7272727	10.173088	4328
[8] {11,SIDEWALK}	=> {NARCOTICS}	0.011053449	0.6929235	9.692611	2164
[9] {11,SIDEWALK,True}	=> {NARCOTICS}	0.011053449	0.9061977	12.675890	2164
[10] {10}	=> {NARCOTICS}	0.009735616	0.2110275	2.951851	1906
lhs <fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1] {11,SIDEWALK,True}	=> {NARCOTICS}	0.011053449	0.9061977	12.675890	2164
[2] {11,ALLEY,True}	=> {NARCOTICS}	0.001522148	0.8612717	12.047465	298
[3] {10,SIDEWALK,True}	=> {NARCOTICS}	0.004142489	0.8069652	11.287826	811
[4] {15,SIDEWALK,True}	=> {NARCOTICS}	0.001935886	0.7641129	10.688409	379
[5] {True,VACANT LOT/LAND}	=> {NARCOTICS}	0.001072654	0.7446809	10.416593	210
[6] {11,True}	=> {NARCOTICS}	0.022106898	0.7272727	10.173088	4328
[7] {11,SIDEWALK}	=> {NARCOTICS}	0.011053449	0.6929235	9.692611	2164
[8] {10,STREET,True}	=> {NARCOTICS}	0.002875736	0.6882641	9.627435	563
[9] {11,STREET,True}	=> {NARCOTICS}	0.004908671	0.6664355	9.322098	961
[10] {SIDEWALK,True}	=> {NARCOTICS}	0.022193732	0.6482172	9.067260	4345

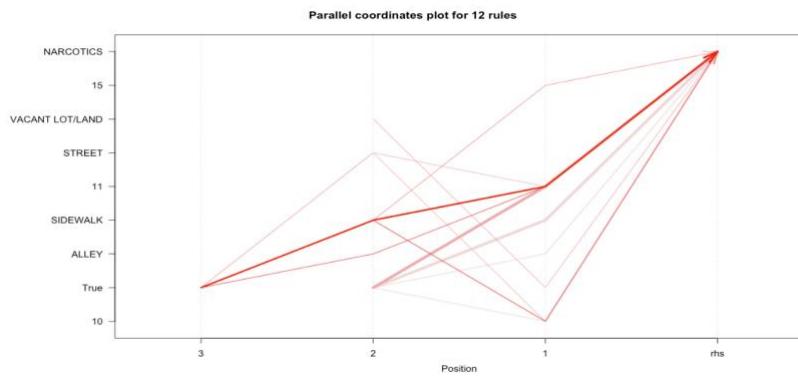
En este caso se ve claramente como si hay varias tendencias bastante claras. La primera es que se ve claramente como hay 3 distritos claramente mas afectados por este delito, que son el 11, 10 y 15. Además se aprecia claramente como el gran numero de arrestos por este crimen en estos distritos, puede hacer el porcentaje de arrestos general en estos barrios sea de los mayores.

Otro detalle a mencionar es que nos encontramos aquí con el primer caso de relación y posible predicción de crímenes futuros, que es el objetivo de este estudio, y es que si un crimen se da en estos barrios, se puede clasificar como NARCOTICS.

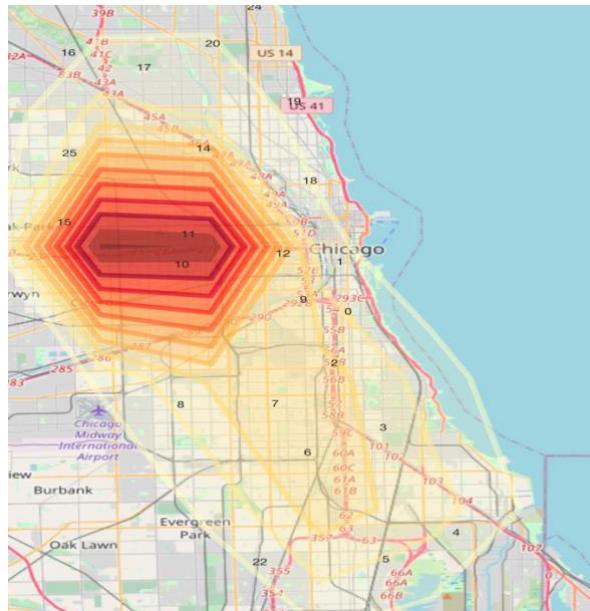
Estas tendencias se pueden ver confirmadas claramente en los siguientes gráficos, siempre recordando que cuanto mas intenso sea el color rojo, mayor es la confianza de esa regla.

Viendo las imágenes parece que a pesar de que los distritos 11, 10 y 15 son los mas afectados, el 11 y el 10 parecen que tienen mas delitos que el 15.





En el mapa, si se confirma claramente tanto lo que se supone del análisis de las reglas de asociación, como lo que se extrae de la exploración de datos, mostrando claramente que los distritos mas afectados son el 10 y el 11.



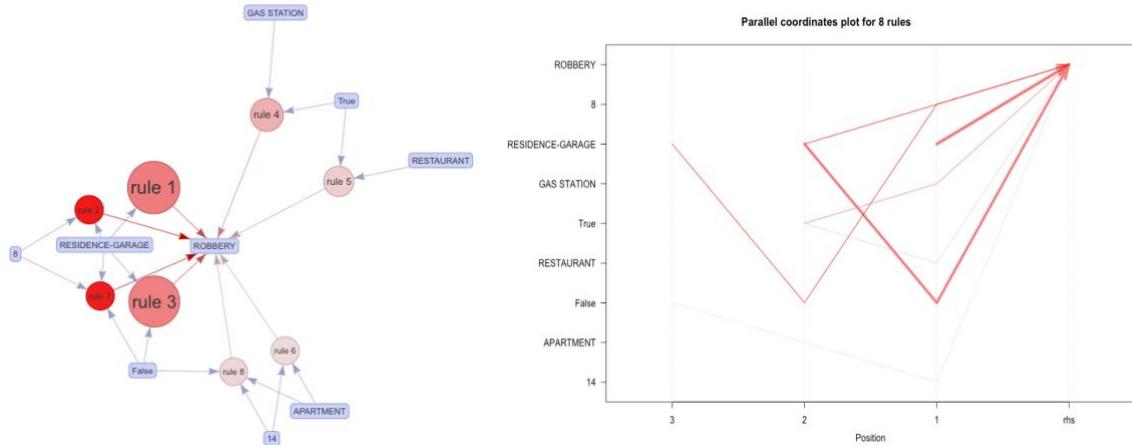
#### - Robbery:

Viendo las siguientes tablas, en este caso no podemos afirmar nada, solo indicar que parece que los distritos 8 y 14 están mas afectados, pero el numero total de veces que ocurren esas relaciones, es demasiado pequeño como para que sean un hecho.

<b>lhs &lt;fctr&gt;</b>	<b>&lt;fctr&gt;</b>	<b>rhs &lt;fctr&gt;</b>	<b>support &lt;dbl&gt;</b>	<b>confidence &lt;dbl&gt;</b>	<b>lift &lt;dbl&gt;</b>	<b>count &lt;int&gt;</b>
[1] {RESIDENCE-GARAGE}	=>	{ROBBERY}	0.007743544	0.4841904	3.999530	1516
[2] {False,RESIDENCE-GARAGE}	=>	{ROBBERY}	0.007340021	0.4804413	3.968562	1437
[3] {GAS STATION,True}	=>	{ROBBERY}	0.002472213	0.4230769	3.494718	484
[4] {RESTAURANT,True}	=>	{ROBBERY}	0.001767326	0.3752711	3.099831	346
[5] {8,RESIDENCE-GARAGE}	=>	{ROBBERY}	0.001282077	0.5691610	4.701408	251
[6] {8,False,RESIDENCE-GARAGE}	=>	{ROBBERY}	0.001220783	0.5663507	4.678194	239
[7] {14,APARTMENT}	=>	{ROBBERY}	0.001159488	0.3552426	2.934390	227
[8] {14,APARTMENT,False}	=>	{ROBBERY}	0.001082870	0.3623932	2.993455	212

<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{8,RESIDENCE-GARAGE}	=> {ROBBERY}	0.001282077	0.5691610	4.701408	251
[2]	{8,False,RESIDENCE-GARAGE}	=> {ROBBERY}	0.001220783	0.5663507	4.678194	239
[3]	{RESIDENCE-GARAGE}	=> {ROBBERY}	0.007743544	0.4841904	3.999530	1516
[4]	{False,RESIDENCE-GARAGE}	=> {ROBBERY}	0.007340021	0.4804413	3.968562	1437
[5]	{GAS STATION,True}	=> {ROBBERY}	0.002472213	0.4230769	3.494718	484
[6]	{RESTAURANT,True}	=> {ROBBERY}	0.001767326	0.3752711	3.099831	346
[7]	{14,APARTMENT,False}	=> {ROBBERY}	0.001082870	0.3623932	2.993455	212
[8]	{14,APARTMENT}	=> {ROBBERY}	0.001159488	0.3552426	2.934390	227

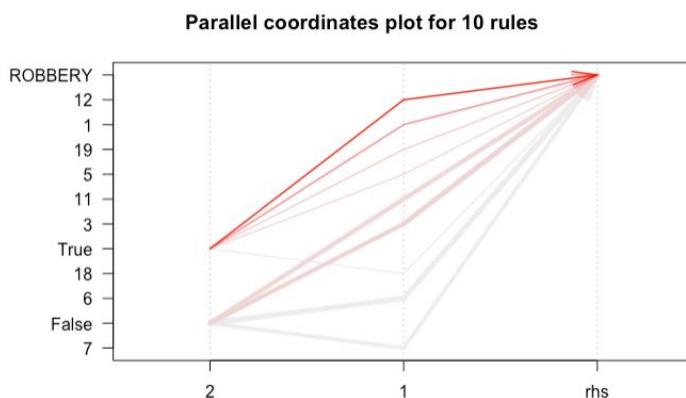
Viendo las siguientes graficas, parece que ni siquiera el 8 podría considerarse, si no solo el 14.



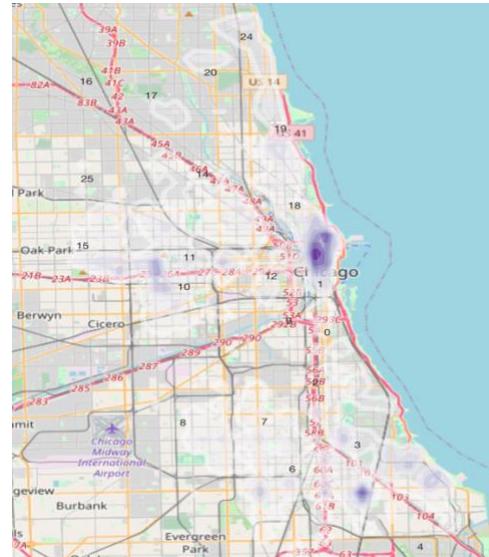
Si limpiamos los datos y nos quedamos con distrito y arresto, el resultado es bastante mas esclarecedor. Se ve como cuantos mas casos hay en un barrio, mas probable es que no se produzca un arresto.

<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{8,False}	=> {ROBBERY}	0.007539229	0.1513536	1.250217	1476
[2]	{6,False}	=> {ROBBERY}	0.006793478	0.1535442	1.268312	1330
[3]	{3}	=> {ROBBERY}	0.006650458	0.1569999	1.296857	1302
[4]	{11,False}	=> {ROBBERY}	0.005909815	0.1591472	1.314594	1157
[5]	{3,False}	=> {ROBBERY}	0.005567587	0.1602470	1.323679	1090
[6]	{7,False}	=> {ROBBERY}	0.005112986	0.1526380	1.260827	1001
[7]	{1,True}	=> {ROBBERY}	0.002027828	0.1781867	1.471865	397
[8]	{12,True}	=> {ROBBERY}	0.001746894	0.2039356	1.684557	342
[9]	{18,True}	=> {ROBBERY}	0.001593658	0.1539221	1.271433	312
[10]	{5,True}	=> {ROBBERY}	0.001419990	0.1607866	1.328136	278

En la siguiente gráfica se diferencian claramente cuales son los barrios donde la probabilidad de arresto es mayor, además se puede asegurar ese hecho con mas confianza.



En este caso el mapa no aporta mucha información nueva ya que lo interesante en este caso es que hay barrios con mas probabilidad de arresto que otros, y eso en el mapa no nos lo indica, ya que este crimen esta bastante repartido, mayormente por el centro-sur de la ciudad.

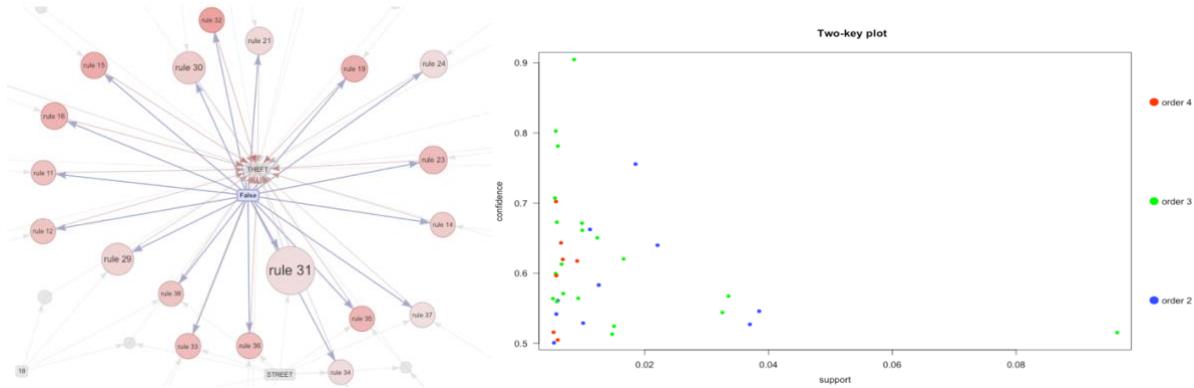


#### - Theft:

Como ocurría con narcóticos, en este caso se ve claramente como los distritos 18 y 1 son por mucha diferencia los mas afectados, además de mostrar una tasa de arresto muy baja para el alto número de delitos que hay de este tipo, el que mas (más de 60000) con diferencia.

<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {False,STREET}	=>	{THEFT}	0.096309047	0.5153188	1.469051	18855
[2] {1}	=>	{THEFT}	0.038482756	0.5458234	1.556012	7534
[3] {18}	=>	{THEFT}	0.036986147	0.5270398	1.502464	7241
[4] {1,False}	=>	{THEFT}	0.033533222	0.5671706	1.616868	6565
[5] {18,False}	=>	{THEFT}	0.032537185	0.5438866	1.550491	6370
[6] {SMALL RETAIL STORE}	=>	{THEFT}	0.022076250	0.6399171	1.824251	4322
[7] {DEPARTMENT STORE}	=>	{THEFT}	0.018521167	0.7557316	2.154410	3626
[8] {False,SMALL RETAIL STORE}	=>	{THEFT}	0.016585281	0.6203668	1.768517	3247
[9] {False,RESTAURANT}	=>	{THEFT}	0.015042702	0.5243012	1.494657	2945
[10] {False,PARKING LOT/GARAGE(NON.RESID.)}	=>	{THEFT}	0.014741337	0.5130667	1.462630	2886
<b>lhs</b> <fctr>		<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {DEPARTMENT STORE,True}	=>	{THEFT}	0.008601667	0.9048898	2.579625	1684
[2] {1,DEPARTMENT STORE}	=>	{THEFT}	0.005633990	0.8027656	2.288493	1103
[3] {GROCERY FOOD STORE,True}	=>	{THEFT}	0.005971110	0.7814171	2.227633	1169
[4] {DEPARTMENT STORE}	=>	{THEFT}	0.018521167	0.7557316	2.154410	3626
[5] {SMALL RETAIL STORE,True}	=>	{THEFT}	0.005490969	0.7072368	2.016163	1075
[6] {1,False,RESTAURANT}	=>	{THEFT}	0.005674853	0.7022756	2.002020	1111
[7] {1,RESTAURANT}	=>	{THEFT}	0.005807658	0.6727811	1.917938	1137
[8] {False,VEHICLE NON-COMMERCIAL}	=>	{THEFT}	0.009863313	0.6714186	1.914054	1931
[9] {GROCERY FOOD STORE}	=>	{THEFT}	0.011165822	0.6624242	1.888413	2186
[10] {DEPARTMENT STORE,False}	=>	{THEFT}	0.009919500	0.6612189	1.884977	1942

En la siguiente imagen se ve el alcance de la baja tasa de arrestos de este crimen. Además, como se ve en la segunda grafica, la confianza con las reglas inferidas en este crimen es la mas alta con diferencia, solo comparable a los crímenes de narcóticos.

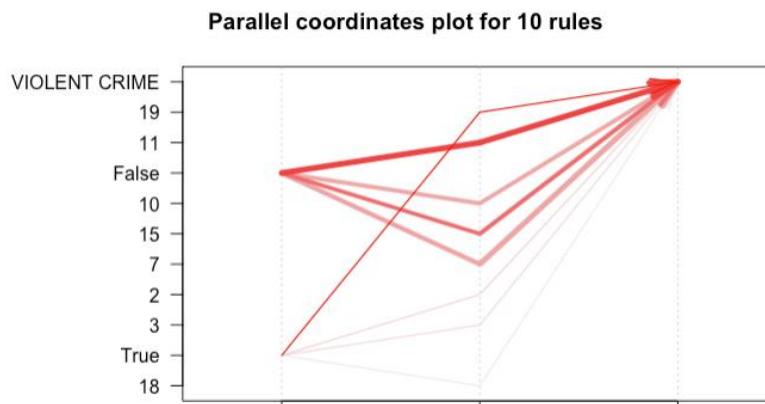


### - Violent Crimes:

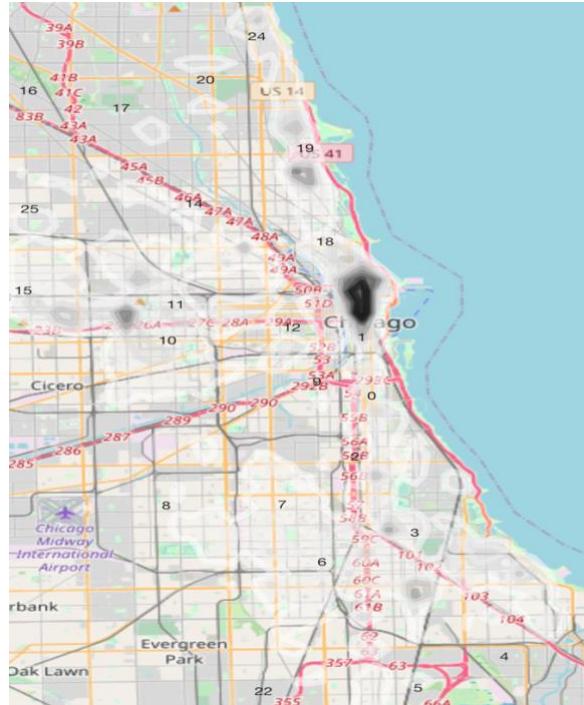
De las tablas no se puede inferir mas que el número de no arrestos es significativamente superior al número de arrestos. En cuanto a distritos ni siquiera sale ninguna regla.

<b>lhs</b> <code>&lt;fctr&gt;</code>	<code>=&gt;</code>	<b>rhs</b> <code>&lt;fctr&gt;</code>	<b>support</b> <code>&lt;dbl&gt;</code>	<b>confidence</b> <code>&lt;dbl&gt;</code>	<b>lift</b> <code>&lt;dbl&gt;</code>	<b>count</b> <code>&lt;int&gt;</code>
[1] {}	=>	{VIOLENT CRIME}	0.13649783	0.1364978	1.0000000	26723
[2] {False}	=>	{VIOLENT CRIME}	0.10873652	0.1364817	0.9998815	21288
[3] {True}	=>	{VIOLENT CRIME}	0.02776132	0.1365647	1.0004895	5435
[4] {SIDEWALK}	=>	{VIOLENT CRIME}	0.02316934	0.2746760	2.0123106	4536
[5] {False,STREET}	=>	{VIOLENT CRIME}	0.01924649	0.1029818	0.7544572	3768
[6] {False,SIDEWALK}	=>	{VIOLENT CRIME}	0.01908814	0.3808990	2.7905131	3737
[7] {APARTMENT}	=>	{VIOLENT CRIME}	0.01552795	0.1742520	1.2765915	3040
[8] {RESIDENCE}	=>	{VIOLENT CRIME}	0.01359717	0.1175380	0.8610977	2662
[9] {APARTMENT,False}	=>	{VIOLENT CRIME}	0.01192179	0.1512638	1.1081771	2334
<b>lhs</b> <code>&lt;fctr&gt;</code>	<code>=&gt;</code>	<b>rhs</b> <code>&lt;fctr&gt;</code>	<b>support</b> <code>&lt;dbl&gt;</code>	<b>confidence</b> <code>&lt;dbl&gt;</code>	<b>lift</b> <code>&lt;dbl&gt;</code>	<b>count</b> <code>&lt;int&gt;</code>
[1] {False,SIDEWALK}	=>	{VIOLENT CRIME}	0.01908814	0.3808990	2.7905131	3737
[2] {SIDEWALK}	=>	{VIOLENT CRIME}	0.02316934	0.2746760	2.0123106	4536
[3] {APARTMENT}	=>	{VIOLENT CRIME}	0.01552795	0.1742520	1.2765915	3040
[4] {APARTMENT,False}	=>	{VIOLENT CRIME}	0.01192179	0.1512638	1.1081771	2334
[5] {True}	=>	{VIOLENT CRIME}	0.02776132	0.1365647	1.0004895	5435
[6] {}	=>	{VIOLENT CRIME}	0.13649783	0.1364978	1.0000000	26723
[7] {False}	=>	{VIOLENT CRIME}	0.10873652	0.1364817	0.9998815	21288
[8] {RESIDENCE}	=>	{VIOLENT CRIME}	0.01359717	0.1175380	0.8610977	2662
[9] {False,STREET}	=>	{VIOLENT CRIME}	0.01924649	0.1029818	0.7544572	3768

Mediante la siguiente gráfica, después de haber limpiado los datos, no se puede deducir que distritos son los mas afectados, ya que su número de casos es muy bajo. Además, cabe destacar que, a pesar de la baja tasa de arrestos general con este crimen, en los barrios 19 y 18, entre otros, la tasa es bastante mayor. También se ve que el índice de arrestos es bastante menor en el oeste y en el sur que en el resto de los distritos.



En el mapa, si lo comparamos con la exploración de datos, la densidad en el centro de la ciudad es bastante engañosa. Parece que se reparte mas en zonas cercanas a la costa y en el oeste.



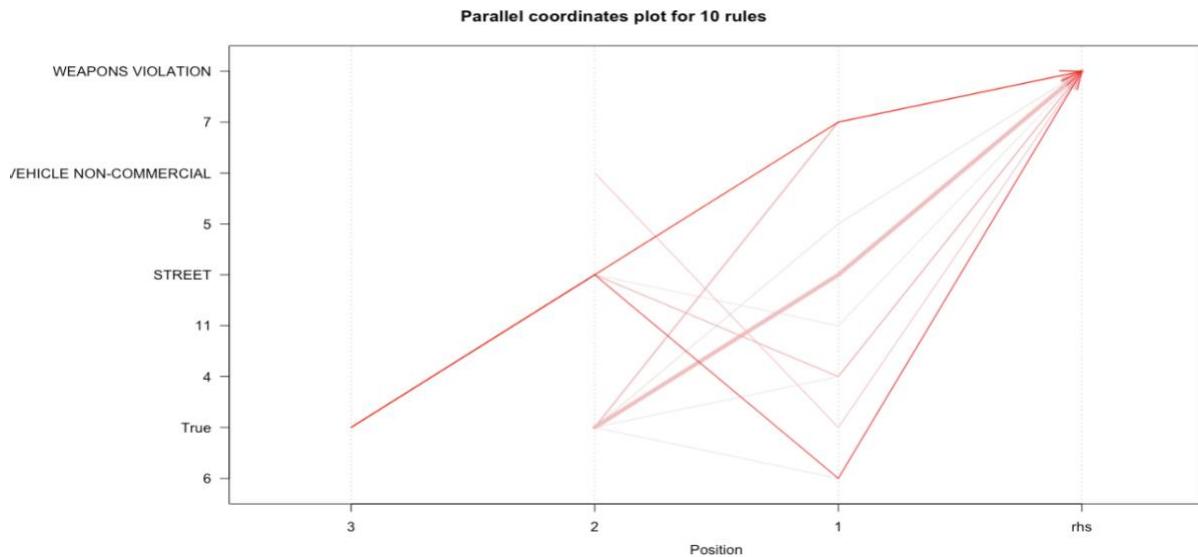
#### - Weapons Violation:

Como ha ocurrido con un par más de crímenes, se ven claramente los distritos mas afectados por este crimen, que son el 4,5,6,7,10. Como se menciona en el apartado 2, estos distritos corresponden al sur, suroeste de la ciudad, que, como se menciona anteriormente, coincide con la zona mas peligrosa y pobre de la ciudad. Además, la tasa de arrestos es bastante alta.

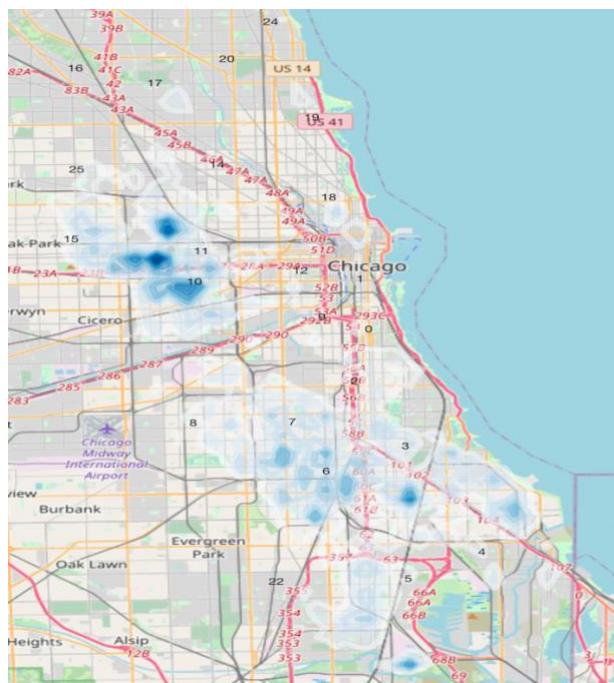
<b>lhs</b> <fctr>	<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {True}	=> {WEAPONS VIOLATION}	0.021499060	0.1057591	3.283916	4209
[2] {STREET,True}	=> {WEAPONS VIOLATION}	0.011155606	0.2223353	6.903715	2184
[3] {ALLEY}	=> {WEAPONS VIOLATION}	0.003197532	0.1478507	4.590900	626
[4] {7,True}	=> {WEAPONS VIOLATION}	0.002712283	0.2352681	7.305288	531
[5] {6,True}	=> {WEAPONS VIOLATION}	0.002410919	0.1683310	5.226830	472
[6] {ALLEY,False}	=> {WEAPONS VIOLATION}	0.002140201	0.1525300	4.736197	419
[7] {4,True}	=> {WEAPONS VIOLATION}	0.001992073	0.1744966	5.418280	390
[8] {6,STREET}	=> {WEAPONS VIOLATION}	0.001951210	0.1301977	4.042757	382
[9] {7,STREET}	=> {WEAPONS VIOLATION}	0.001854160	0.1507475	4.680848	363
[10] {10,True}	=> {WEAPONS VIOLATION}	0.001782650	0.1111465	3.451200	349
<b>lhs</b> <fctr>	<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1] {7,STREET,True}	=> {WEAPONS VIOLATION}	0.001614090	0.3722026	11.557230	316
[2] {6,STREET,True}	=> {WEAPONS VIOLATION}	0.001583442	0.3329753	10.339187	310
[3] {4,STREET,True}	=> {WEAPONS VIOLATION}	0.001215675	0.2650334	8.229529	238
[4] {7,True}	=> {WEAPONS VIOLATION}	0.002712283	0.2352681	7.305288	531
[5] {True,VEHICLE NON-COMMERCIAL}	=> {WEAPONS VIOLATION}	0.001006252	0.2328605	7.230532	197
[6] {STREET,True}	=> {WEAPONS VIOLATION}	0.011155606	0.2223353	6.903715	2184
[7] {5,True}	=> {WEAPONS VIOLATION}	0.001619197	0.1833430	5.692967	317
[8] {11,STREET,True}	=> {WEAPONS VIOLATION}	0.001317833	0.1789182	5.555572	258
[9] {4,True}	=> {WEAPONS VIOLATION}	0.001992073	0.1744966	5.418280	390
[10] {6,True}	=> {WEAPONS VIOLATION}	0.002410919	0.1683310	5.226830	472

Viendo la siguiente gráfica, además de confirmar lo propuesto anteriormente, parece encontramos la primera distinción en porcentajes de arresto entre distritos, como puede ser

que las dos reglas con mayor confianza son la de arresto TRUE y distrito tanto 7 o 6, indicando mas proclividad para arrestar en estos dos distritos que en el resto.



En este caso, el mapa clarifica el reparto de crímenes con un número mayor en el oeste y sur de la ciudad, estando mas concentrados en los distritos 11 y 10 en el oeste y mayor numero pero menos densidad en el sur.



#### - Arrestos

Si miramos las tablas de reglas generadas si buscamos información sobre arresto *True* o *False*, confirmamos varios hechos que se intuyen de los anteriores análisis.

En cuanto a mayor tasa de arresto, hay dos crímenes que son los únicos con una tasa de arresto de mas del 50%, que son WEAPONS VIOLATION y NARCOTICS.

	<b>lhs</b> <fctr>	<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{NARCOTICS}	=> {True}	0.07144900	0.9994284	4.916430	13988
[2]	{WEAPONS VIOLATION}	=> {True}	0.02149906	0.6675654	3.283916	4209

En cuanto a crímenes donde no se cometen arrestos, destacan THEFT, CRIMINAL DAMAGE, DECEPTIVE PRACTICE y ASSAULT, dejando ROBBERY y VIOLENT CRIMES en una mezcla entre arrestos y no arrestos.

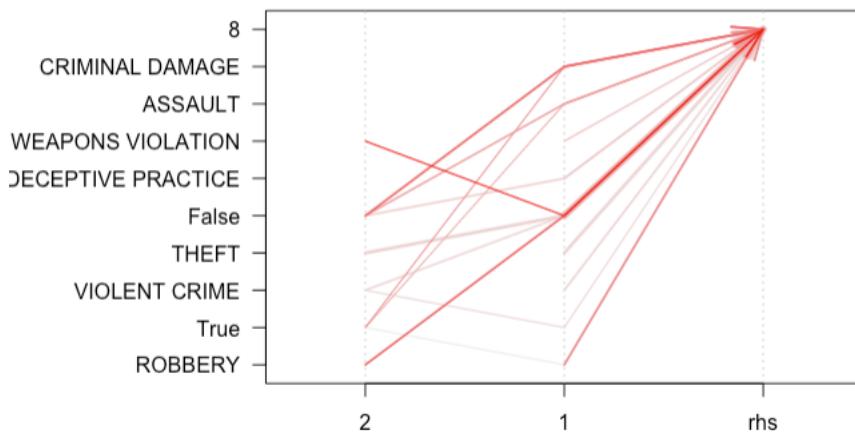
Además, se ve en que distritos se producen menos arrestos, lógicamente donde esos crímenes son mas pronunciados, donde se ve una tendencia, ya que son los distritos mas ricos de la ciudad y teóricamente mas seguros.

Es importante hablar ahora del distrito 8. Es de los que mas crímenes hay, pero prácticamente no sale en ninguno de los análisis hechos, solo para indicar que en el 85% de los crímenes del distrito no hay arrestos, lo cual no he encontrado ninguna explicación aún, por lo que parece que habrá que monitorizar la evolución de este barrio a lo largo del estudio.

	<b>lhs</b> <fctr>	<b>rhs</b> <fctr>	<b>support</b> <dbl>	<b>confidence</b> <dbl>	<b>lift</b> <dbl>	<b>count</b> <int>
[1]	{THEFT}	=> {False}	0.317965430	0.9064434	1.137731	62250
[2]	{CRIMINAL DAMAGE}	=> {False}	0.113103751	0.9445463	1.185556	22143
[3]	{DECEPTIVE PRACTICE}	=> {False}	0.085863436	0.9538671	1.197255	16810
[4]	{ASSAULT}	=> {False}	0.063628841	0.8137044	1.021329	12457
[5]	{18}	=> {False}	0.059823472	0.8524638	1.069978	11712
[6]	{1}	=> {False}	0.059123692	0.8385858	1.052559	11575
[7]	{8}	=> {False}	0.049812030	0.8490336	1.065673	9752
[8]	{12}	=> {False}	0.046818813	0.8453380	1.061034	9166
[9]	{19}	=> {False}	0.046308025	0.8711444	1.093425	9066
[10]	{25}	=> {False}	0.038385706	0.8092828	1.015779	7515

Para este distrito en concreto he hecho el análisis y el resultado es el siguiente.

Parallel coordinates plot for 19 rules



Parece que en este barrio no hay ninguna tendencia específica de un crimen o varios crímenes, sino que están todos muy repartidos, teniendo en cuenta la proporcionalidad entre ellos (THEFT hay 60000 totales y WEAPON VIOLATION 6300).

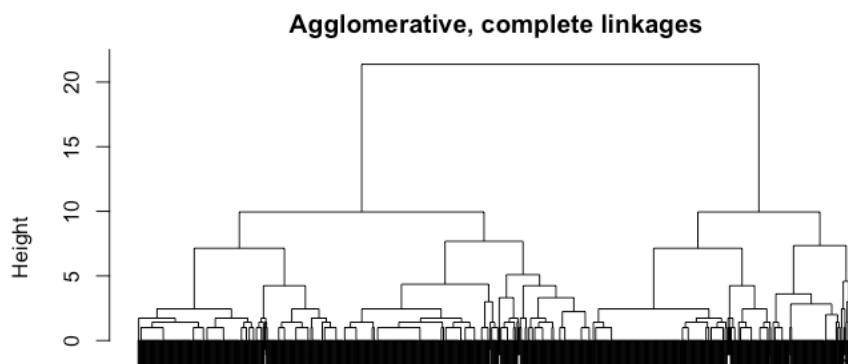
- TABLA RESUMEN

A continuación, se muestra una tabla a modo de resumen para tener en cuenta lo deducido hasta ahora en los siguientes pasos.

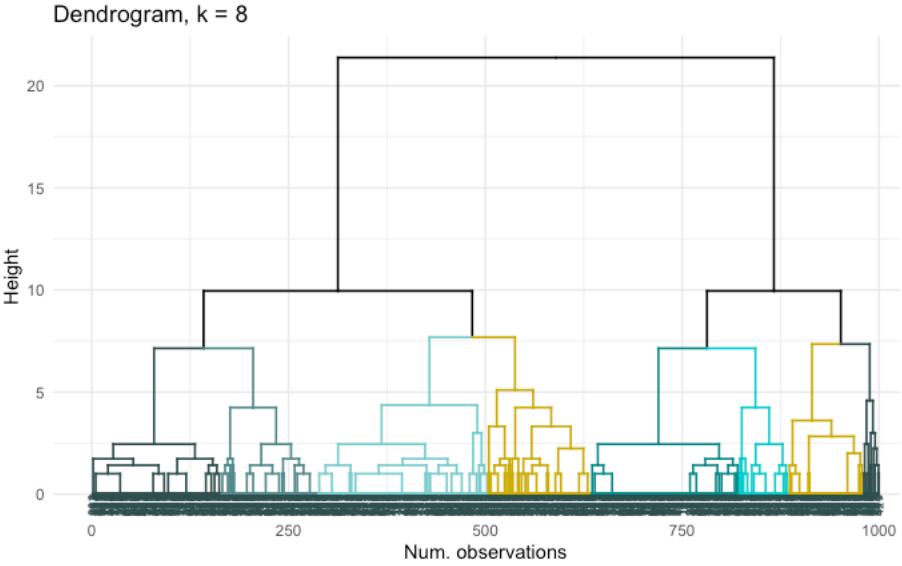
	<b>Arresto</b>	<b>Distrito</b>	<b>Seguridad</b>
<b>Assault</b>	Mayoritariamente No (Menos el distrito 4)	7, y, en menor medida 3,4,5 y 6	Media
<b>Criminal Damage</b>	No	8 y la zona sur y la zona centro	Alta
<b>Deceptive Practice</b>	No	Zona centro (1,18,19)	Alta
<b>Narcotics</b>	Si	10 y 11 (15 en menor medida)	Alta
<b>Robbery</b>	Mayoritariamente No (En los centricos, 1,12,19,18 Si)	Bastante repartido por la ciudad	Media-Baja
<b>Theft</b>	No	1,18,19	Alta
<b>Violent Crime</b>	No (Mayor en los barrios mas centricos)	Bastante repartido por la ciudad	Media-Baja
<b>Weapons Violations</b>	Si	Sur (4,5,6,7) y en menor medida Oeste	Alta

#### 4. Aprendizaje Automático: Clustering y Arboles de Decision

En este apartado voy a tratar de comprobar si a partir de un aprendizaje a partir de clustering y arboles de decisión, podemos comprobar las mismas relaciones que anteriormente, solo que estos algoritmos nos proporcionan una certeza mayor, al ser algoritmos mas complejos.



En la anterior imagen vemos todos los clusters en los que se divide el set, no se aprecia mucho, pero si se puede ver en la siguiente imagen que, sin muchas iteraciones si se dividen fácilmente los clusters en los tipos de crímenes que buscamos.

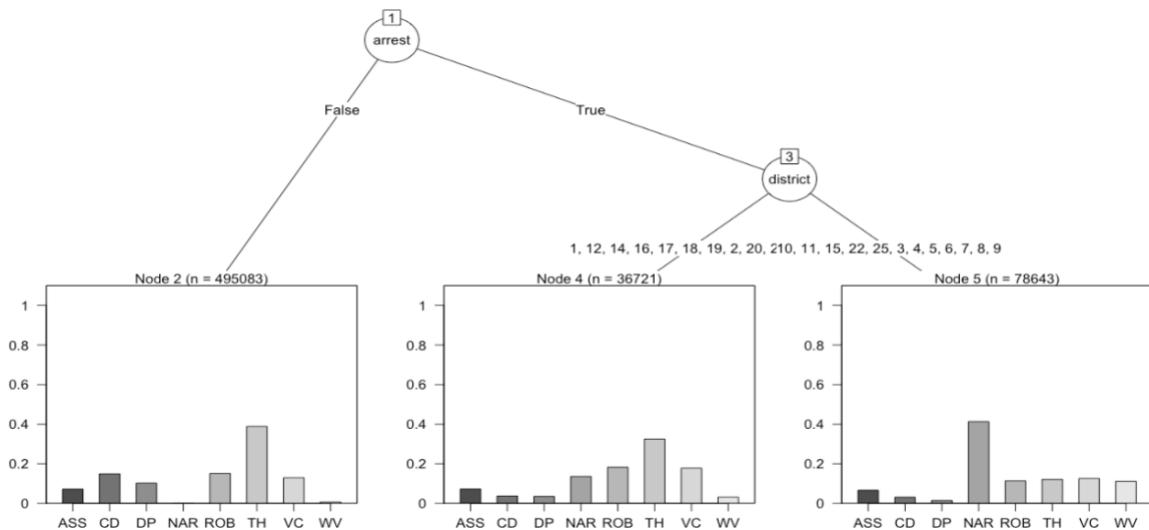


Siendo de izquierda a derecha los tipos, THEFT, ROBBERY, VIOLENT CRIME, CRIMINAL DAMAGE, DECEPTIVE PRACTICE, ASSAULT, NARCOTICS y WEAPONS VIOLATION.

Este método resulta demasiado engorroso y no muestra mucho conocimiento mas que el que ya sabíamos que había 8 tipos de crímenes.

Sin embargo, si hacemos un análisis por cada crimen para buscar los distritos mas reincidentes usando arboles de decisión, si obtenemos conocimiento que nos puede servir.

Empezando por un análisis general, en la siguiente figura se observa como se distribuyen los crímenes según diferentes condiciones.

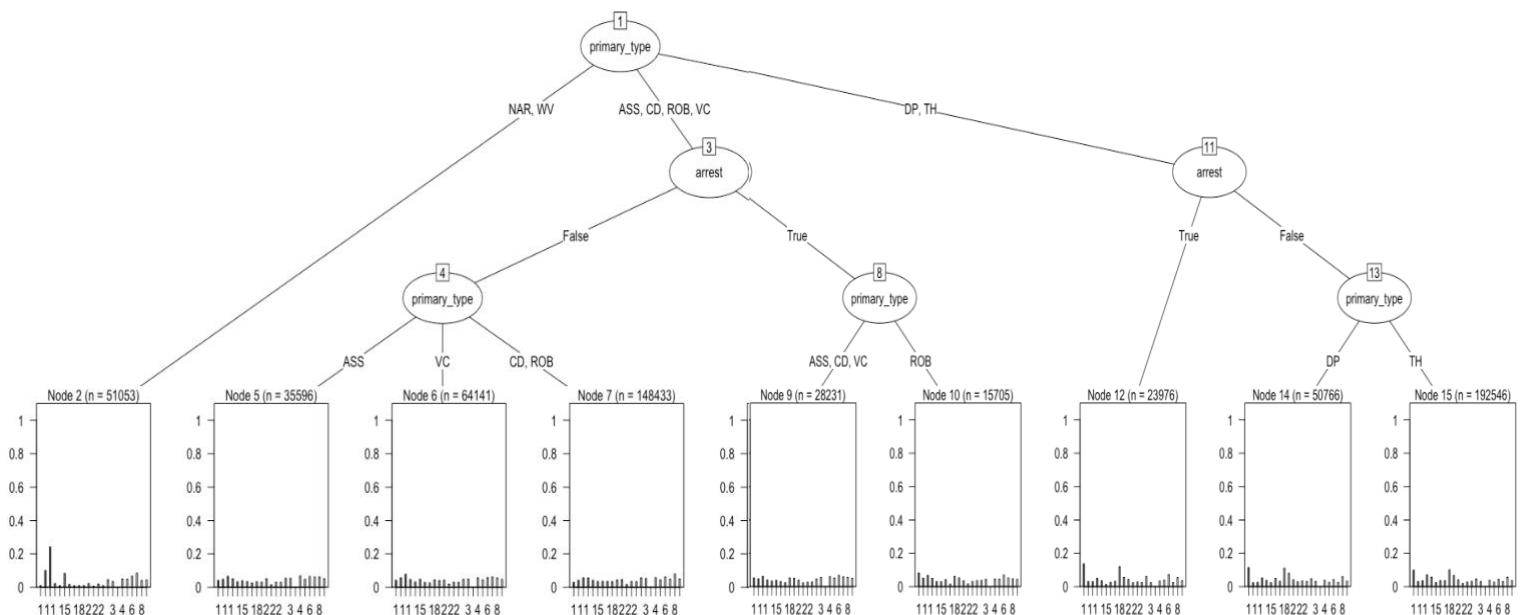


El clasificador funciona razonablemente bien, arrojando las siguientes conclusiones. Usando el primer separador de arrestos, clasifica los eventos en su gran mayoría como THEFT, ahí es donde obtenemos la mayor parte del error, ya que no solo. Son THEFT si no que también hay un considerable número de DECEPTIVE PRACTICE, VIOLENT CRIMES, ROBBERY y

VIOLENT CRIME, y no todos se clasifican. Lo que si hace muy bien es no incluir casi ningún crimen de NARCOTICS o WEAPONS VIOLATION, que es algo que si nos puede interesar.

Usando el separador de distritos, básicamente usa un separador en distritos del sur y oeste (3,4,5,6,7,8, 9,10, 11, 15 y 22) y distritos del centro y norte, que son el resto. Esto nos da resultados muy interesantes ya que, como supuse, cuando hay arrestos, en los barrios del norte y centro, los crímenes suelen ser THEFT, ROBBERY o VIOLENT CRIME. Y en los otros barrios de NARCOTICS y WEAPONS VIOLATION en su mayoría.

El error de este árbol es del 40%, que es bastante alto, muy probablemente debido a que no divide por barrios los crímenes sin arresto, que son la gran mayoría de los casos. Por ello, voy a ver si un árbol que haga otra clasificación ayuda a cerciorar estas conclusiones, ya que este árbol solo me sirve de orientación.



En el árbol anterior vemos como si clasificamos los distritos el árbol es significativamente distinto. No solo tiene mas información, sino que, además, el error baja hasta un 12%, mucho mas aceptable.

El algoritmo ha hecho las siguientes separaciones, la primera ha sido agrupar NARCOTICS y WEAPONS VIOLATION, clasificando la gran mayoría de crímenes en el distrito 11 y, seguidamente en el 10, 5,6,7 y 8. Esta distribución confirma nuestra hipótesis de distribución de crímenes de NARCOTICS, y, aunque haya fallos, también de WEAPONS VIOLATION, ya que, aunque los distritos 10 y 11 no están entre los mas problemáticos, si lo están el resto.

La segunda agrupación que hace el algoritmo es ASSAULT, CRIMINAL DAMAGE, ROBBERY y VIOLENT CRIMES. Si, además el arresto se produce, la distribución es bastante similar por toda la ciudad, pero para ASSAULT, CRIMINAL DAMAGE y VIOLENT CRIMES, el distrito mas afectado es el 6, distrito bastante afectado por ASSAULT y CRIMINAL DAMAGE, no tanto por VIOLENT CRIMES, que esta bastante repartido por la ciudad. Si el arresto es falso, de nuevo

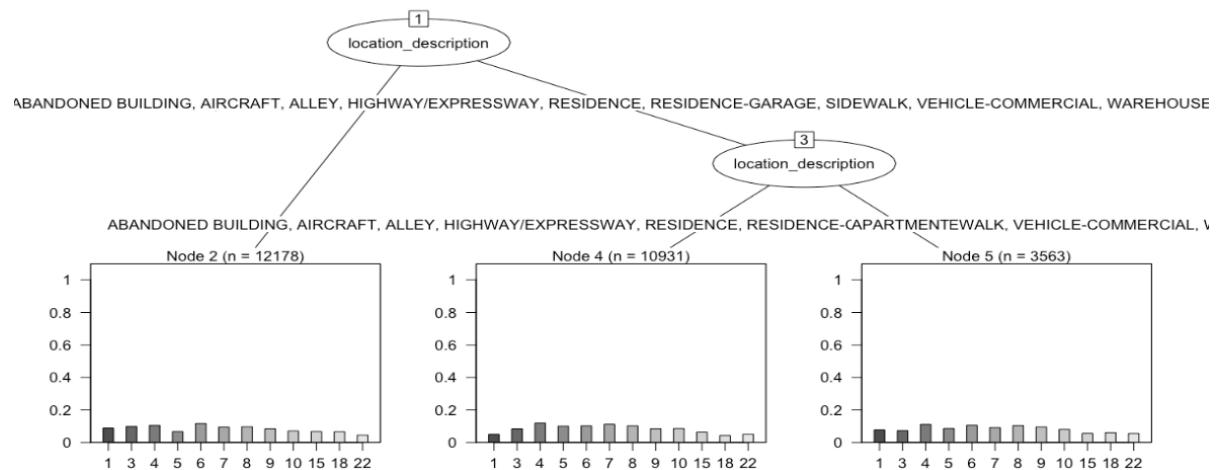
la distribución parece bastante equilibrada, esto se debe a que, aunque para estos cuatro crímenes, hay distritos predominantes, no lo son muy claramente. En este caso, el distrito mas afectado por ASSAULT es el 4, por VIOLENT CRIMES es el 11 y para CRIMINAL DAMAGE y ROBBERY el 8, que recordemos, es de los distritos con mas crimen, pero no he podido encontrar ningún patrón antes con el.

Por ultimo, separa con DECEPTIVE PRACTICE y THEFT. Si hay arresto, para ambos crímenes el distrito mas afectado es el 1, sin embargo, aunque la distribución es similar, si no hay arresto, el distrito mas afectado por THEFT es el 18 y por DECEPTIVE PRACTICE el 1.

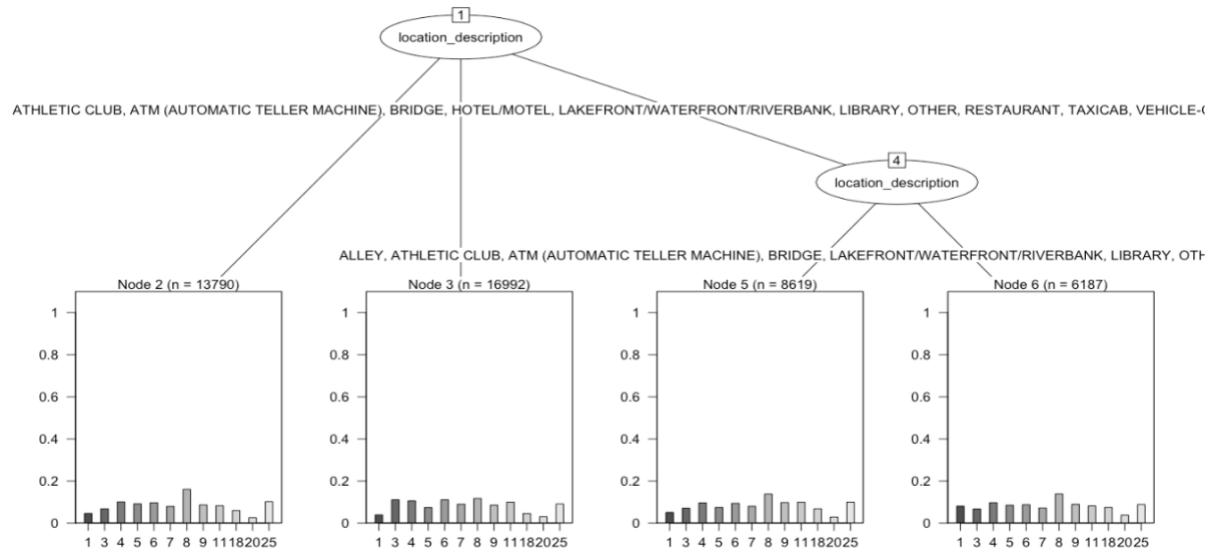
En conclusiones generales, aplicando este método, parece que las hipótesis que había deducido de apartados anteriores se confirman claramente para los crímenes de NARCOTICS, WEAPONS VIOLATION, THEFT y DECEPTIVE PRACTICE. Sin embargo, aunque para el resto, los resultados de los arboles si indican una similitud, no es tan claro como yo desearía o como debería.

Centrándonos en cada crimen de esos crímenes que faltan, voy a intentar conseguir mejores estimaciones con los teniendo en cuenta solo esos crímenes que están dudosos. Todos estos arboles tienen entre un 15% y un 20% de error, que no es óptimo, pero es aceptable.

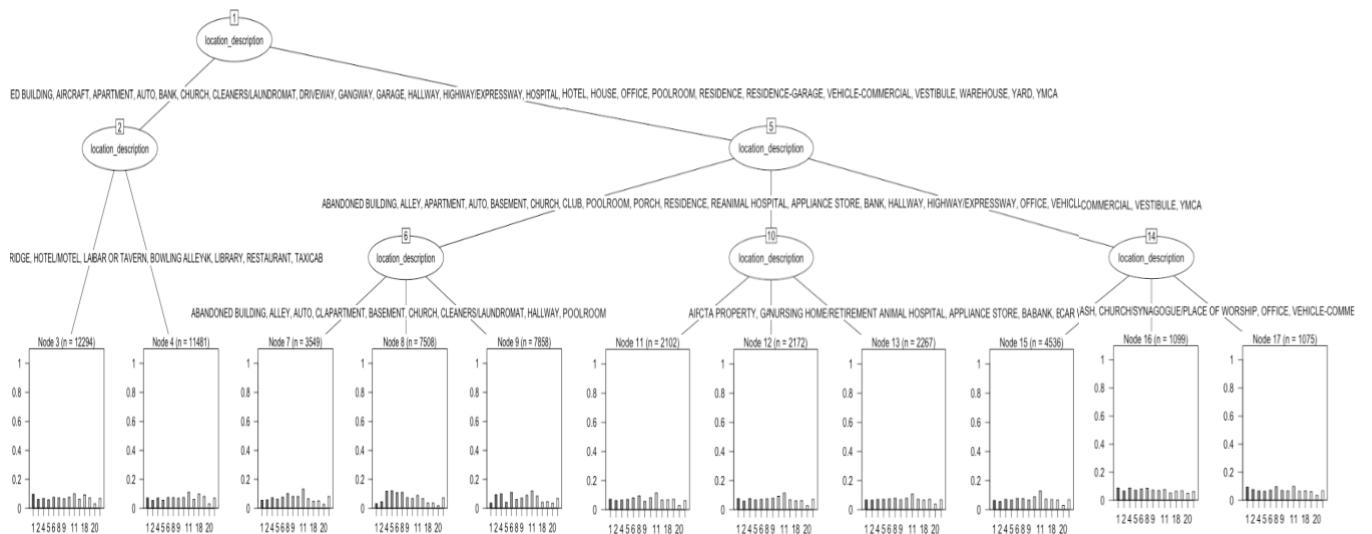
Empezando por ASSAULT, el resultado es el siguiente:



En base a los resultados obtenidos, lo único que podemos asegurar es que la zona norte de la ciudad esta bastante menos afectada que el sur y el oeste, sirviendo el centro de la ciudad (1 y 18) como transición. No se puede asegurar que determinados distritos estén mas afectados que otros dentro de la misma zona.



Aquí, para CRIMINAL DAMAGE, el clasificador si nos permite ver un distrito que está mas afectado que el resto, que es el 8. A partir de ahí, no se puede destacar ninguno por encima del otro. Si parece que podemos quitar el centro como distritos mas afectados. Siguen teniendo este problema, pero hay crímenes mucho mas presentes en esa zona.



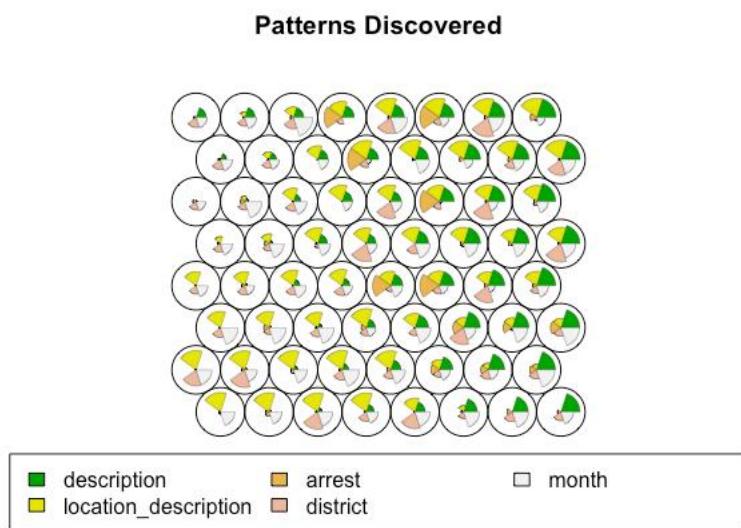
En el anterior árbol, correspondiente a VIOLENT CRIMES, vemos como tampoco podemos asegurar ninguna de las hipótesis mas que en el norte es una zona un poco mas segura.

## 5. Aprendizaje a base de ejemplos: SOM y Redes Neuronales

En la parte final de este estudio, voy a utilizar métodos mas profundos y complejos para seguir confirmando las relaciones ya mostradas a lo largo de este estudio. Para ello, voy a utilizar métodos de mapas auto-organizados, SOM y redes neuronales, ambos métodos caracterizados por su aprendizaje en base a ejemplos.

### - SOM:

Haciendo un primer análisis general, usando las variables *description*, *location\_description*, *arrest*, *district* y *month*, obtenemos los siguientes patrones de clasificación que muestran cuanto se usa cada variable para la clasificación en cada neurona.

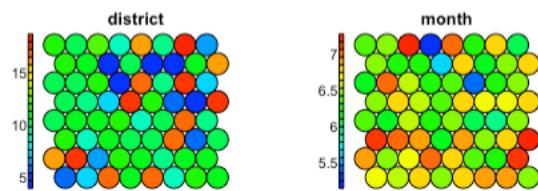
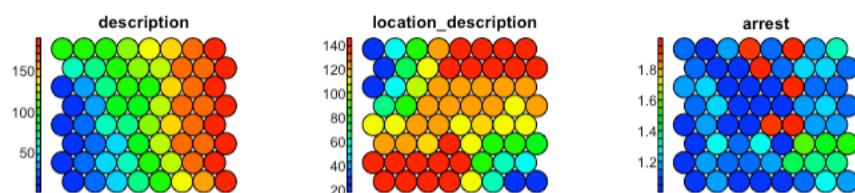
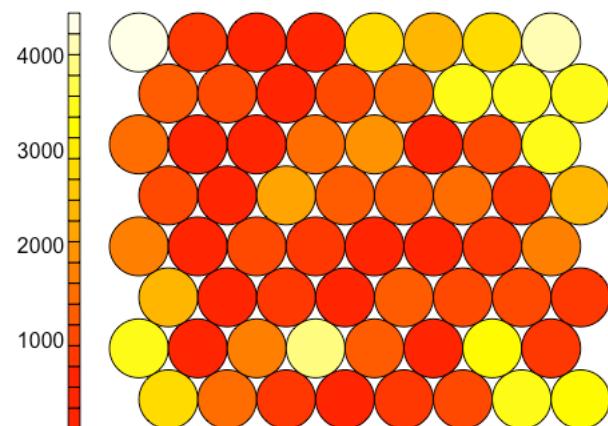


Aquí se ve como se utiliza cada variable en cada neurona, si vemos en la siguiente imagen el numero de ejemplos utilizados por cada neurona, vemos que, lógicamente, en las neuronas que mas ejemplos se han utilizado, se han utilizado mayoritariamente, por no decir en su totalidad, las variables de descripción, localización y distrito, lo que indica que es muy fácil clasificar cada ejemplo con esas variables.

Una ayuda para comprender ese reparto se muestra a continuación del numero de ejemplos, usando una codificación de colores de mas frio (azul oscuro) a mas caliente (Rojo).

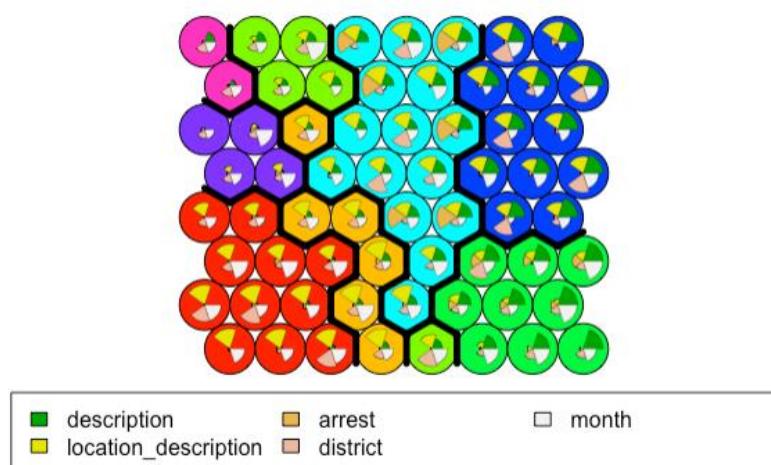
En este análisis se obtiene un 77% de acierto, que, si soy capaz de mantenerlo usando solo distritos y arrestos, seria muy interesante y indicaría que mi hipótesis de reparto de crímenes por distritos es muy válida.

### Examples per Neuron



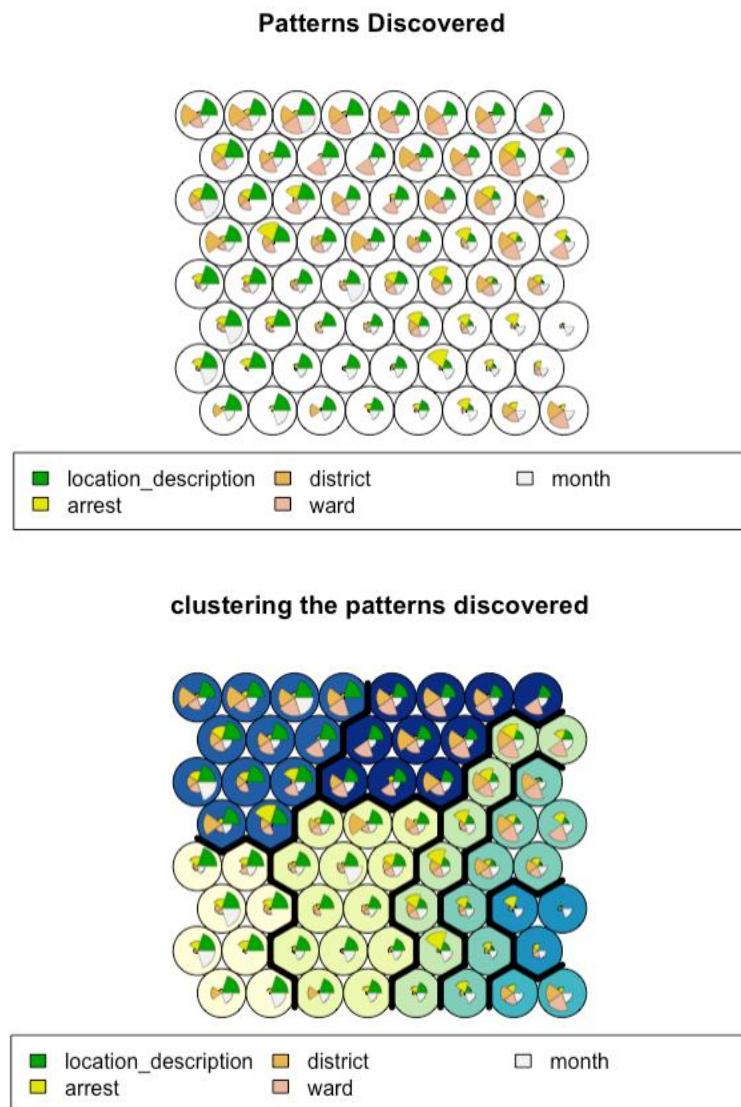
Haciendo un clustering de los grupos formados con este método, se dividen en las 8 clases siguientes.

### clustering the patterns discovered



Ahora, haciendo un análisis solo con las variables que nos interesan, que son arresto y localidad, vamos a ver que resultado obtenemos.

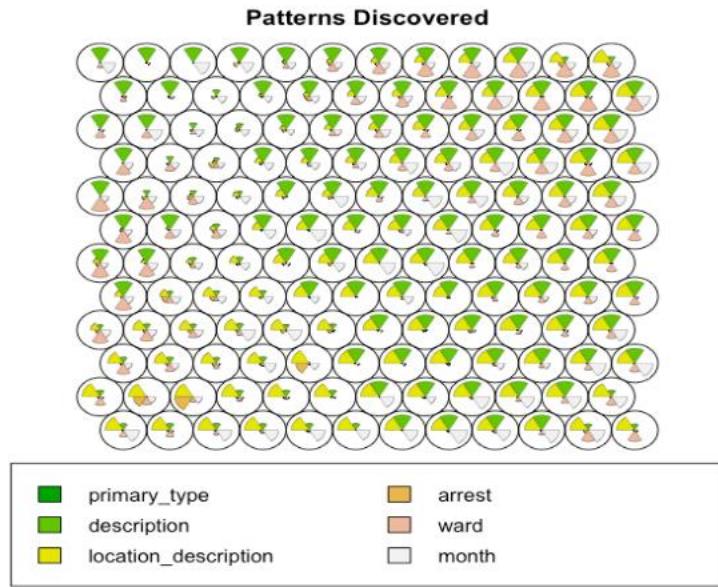
Haciendo ese análisis se obtienen los siguientes patrones y clases.



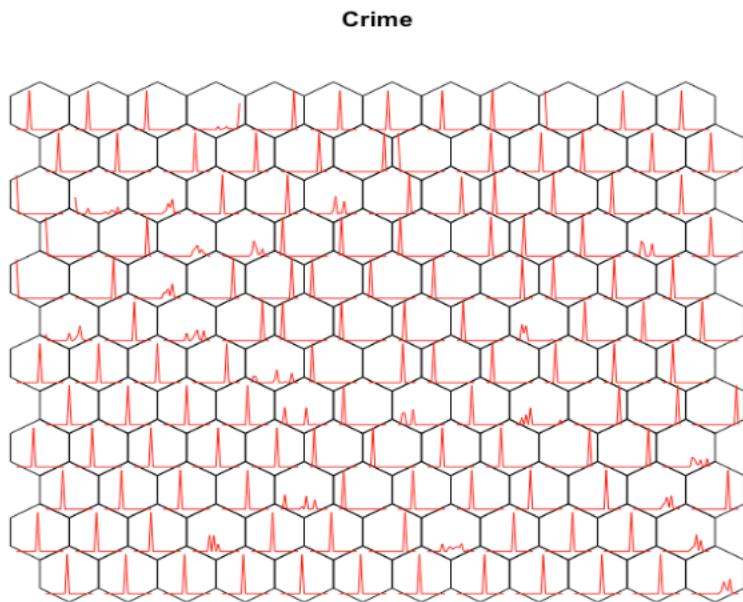
A pesar de que este resultado inicial parezca bueno, el acierto del modelo es de un 40% aproximadamente. Esto me indica que, hay relaciones, pero examinándolas juntas no se ven tan claramente.

Si probamos por crímenes, obtenemos los siguientes resultados.

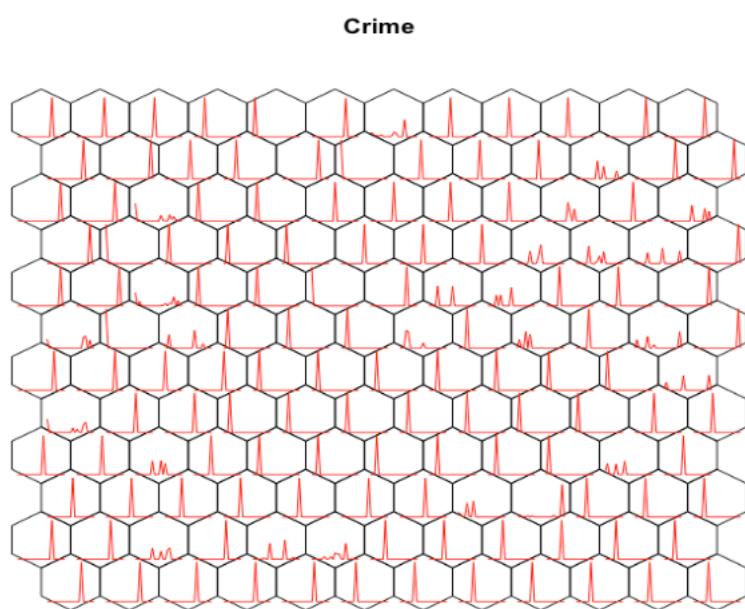
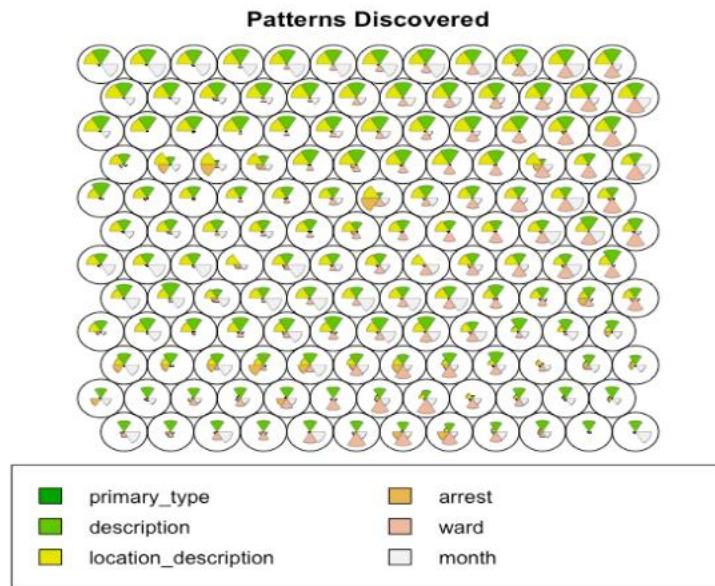
Empezando por ASSAULT, el modelo predice los distritos con una precisión balanceada del 65% con los patrones y grupos siguientes. Los distritos que predice con una mejor precisión son 4,6,7,8, 17. Obtenemos con estos resultados un distrito que no encontrábamos en análisis anteriores, que es el 17. El resto son acorde a lo mostrado anteriormente, afectando mayoritariamente al sur, suroeste de la ciudad.



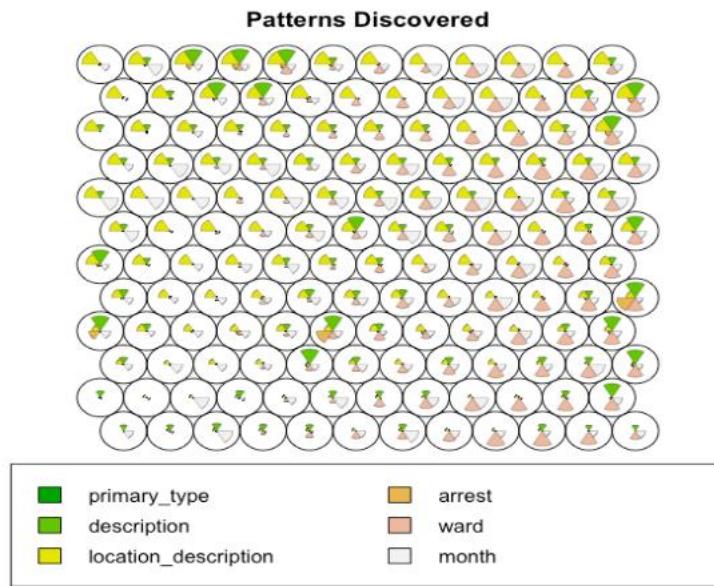
Si miramos la siguiente imagen, hay neuronas que si les cuesta bastante clasificar un distrito, pero la mayoría lo hacen.



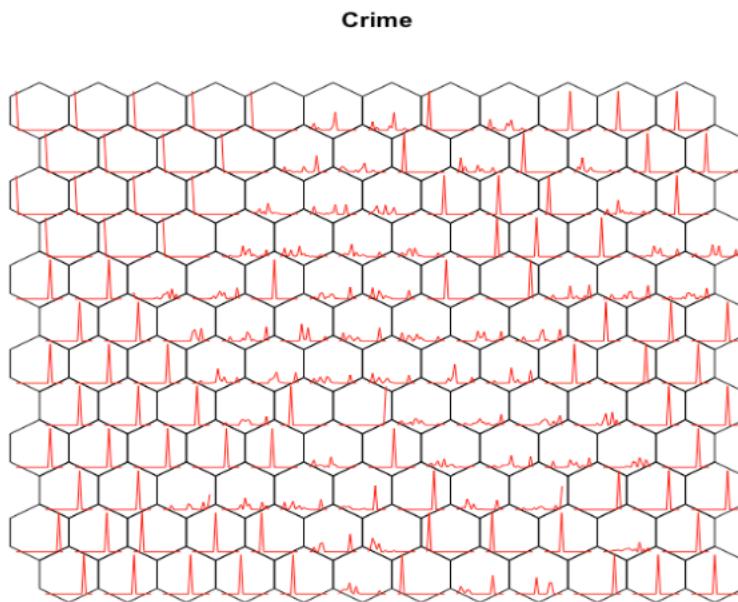
CIMINAL DAMAGE obtiene los siguientes resultados, con un 70% de acierto clasificando los distritos 1, 2, 5, 7, 8, 19 y 24 con al menos un 75%, coincidiendo con las suposiciones de apartados anteriores, añadiendo el distrito 24, que en apartados anteriores he considerado como un outlier, pero cada vez parece mas un foco de crimen en una zona muy poco común, ya que es el distrito mas al norte de la ciudad.



En cuanto a DECEPTIVE PRACTICE, la precisión balanceada es de alrededor del 65%, y la precisión de cada distrito es muy similar entre si, excepto la del distrito 24. Además de la precisión, el numero de casos encontrados es mucho mayor para los distritos 1,2,18 y 8 , junto al 24 que al resto. En análisis anteriores daba por hecho que solo afectaba al centro, pero según este análisis parece que además afecta al distrito 24 y al 8, ambos destacándose como focos de criminalidad frente a los que les rodean (Una diferencia mucho mayor el 24 que el 8, posiblemente debida a su localización geográfica). Su representación es la siguiente.

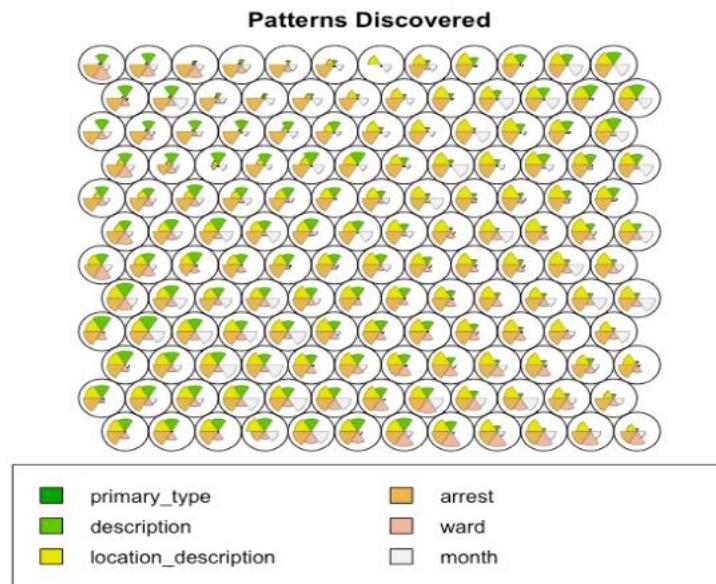


Con el análisis de patrones en este caso, se ve que al modelo le cuesta mucho clasificar muchos de los distritos. Asumo que esto se debe a que, fuera de los distritos mencionados anteriormente, el resto son muy similares, bajando la precisión total del modelo considerablemente. Esto ocurre también en el resto de crímenes, fuera de sus distritos mas afectados.

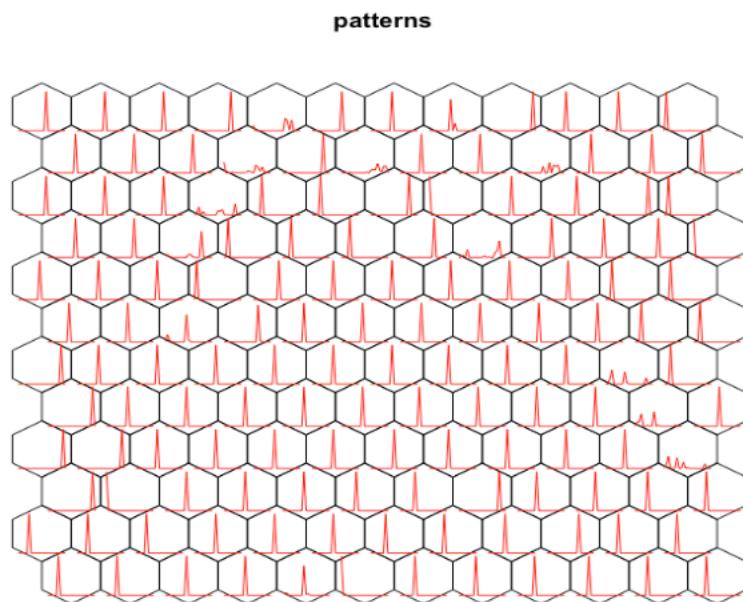


El análisis de NARCOTICS muestra unas conclusiones un poco sorprendentes que, sin embargo, tiene bastante sentido. El porcentaje de acierto es del 75%, el mas alto con diferencia, llegando a mas del 80% en los distritos 10 y 11, probablemente la causa de la subida, pero el distrito con mas acierto es el 18, con un 85%. Esto, en mi opinión tiene una explicación, el distrito 18 esta mayormente afectado por delitos cuya tasa de arrestos es ridículamente baja, como THEFT o DECEPTIVE PRACTICE, por lo que la intromisión de un delito con una tasa de arrestos cercana al 100%, hace que sobresalga. En total, el numero de

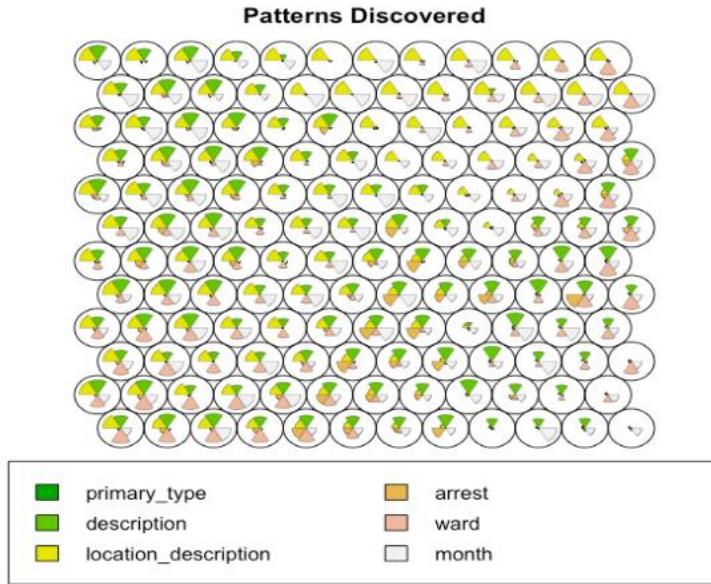
casos es muy pequeño en comparación con otros distritos como el 15 o el 9, pero creo que es un factor digno de mencionar. A continuación, se ven los resultados gráficos.



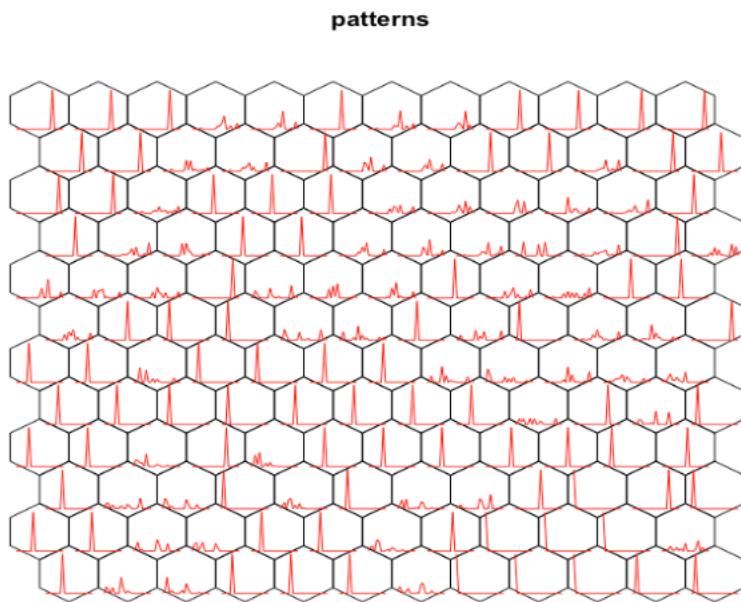
Observando los patrones, claramente al modelo no le cuesta mucho clasificar los distritos en este caso, permitiéndome estar bastante seguro de mis suposiciones anteriores.



ROBBERY es de los que menos porcentaje de acierto tiene, siendo un 60% raspado. Sin embargo, el distrito 24 tiene un 82% de acierto, aunque el crimen esté muy repartido por la ciudad, con pequeños focos en los distritos centrales, donde ya se ha visto que el arresto es mas probable que en el resto, en el distrito 4 y el 24. De todos ellos, solo tiene un porcentaje de clasificación alto el 24.

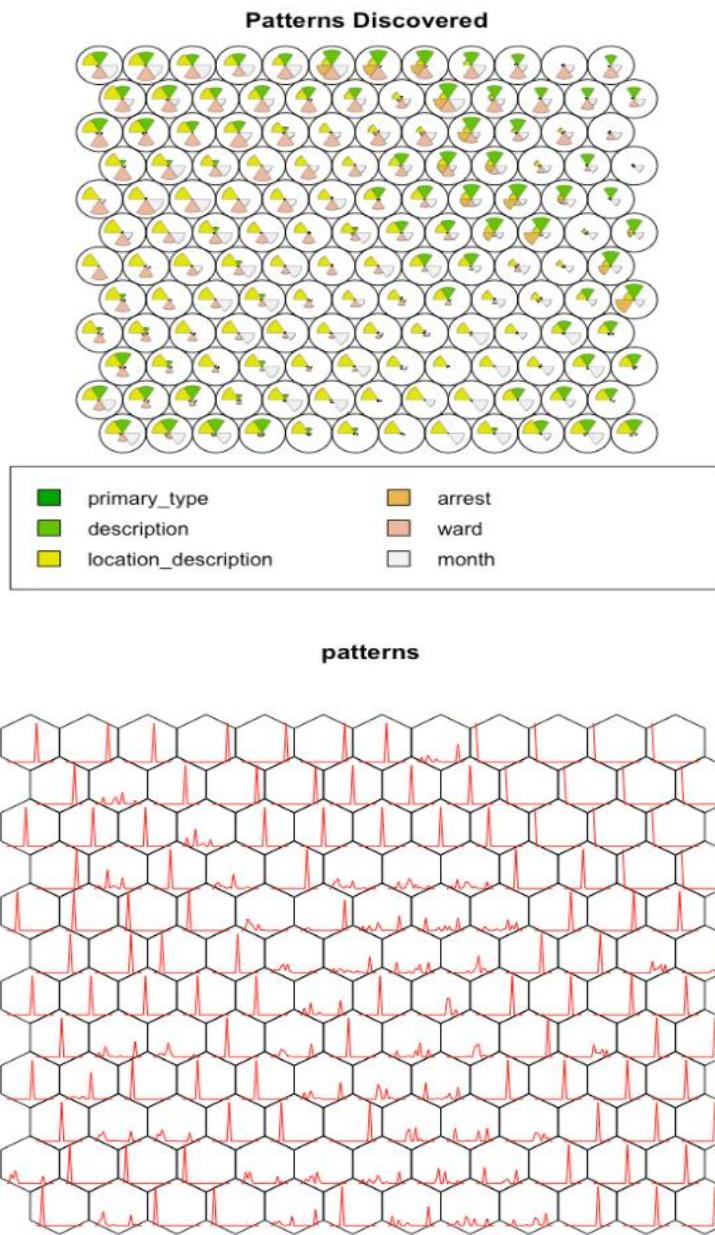


En el análisis de patrones se ve lo que confirman los números, no se consigue clasificar el distrito con mucha facilidad en este crimen.



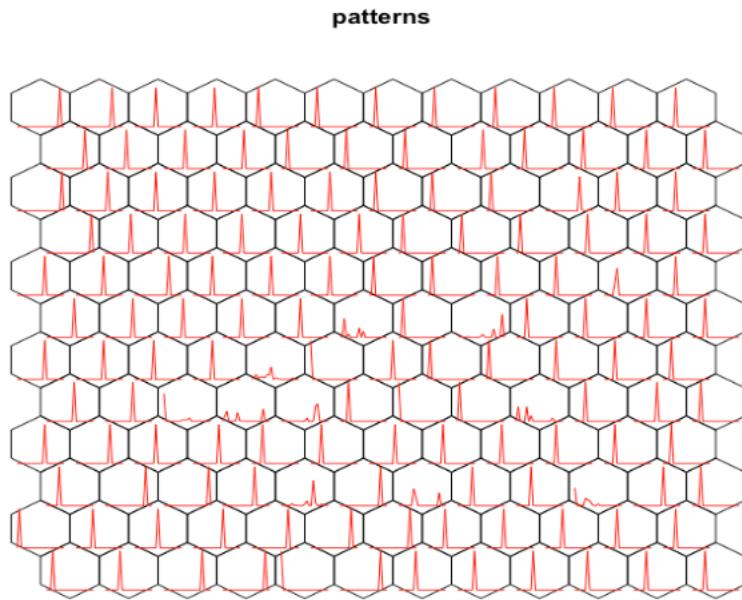
THEFT arroja unos resultados bastante esperables. El porcentaje de acierto general es del 70%, un resultado aceptable, sin embargo, es de un 77% en el distrito 24 y de un 96% en el 16. De nuevo nos encontramos con el distrito 24, que no parece muy predominante en ningún crimen concreto, pero si en general, teniendo además unas características muy particulares, para que la clasificación sea tan fácil. Lo mismo parece ocurrir con el distrito 16, pero al parecer solo en este caso, posiblemente deben ser los únicos distritos con una tasa de arrestos mayor (Esto se comprueba muy claramente en los diagramas de correlaciones expuestos en la exploración de datos).

Los distritos mas afectados, de todas maneras, siguen siendo los distritos 1, 18 y 19, el centro de la ciudad, como se ve en los análisis de patrones siguientes.



VIOLENT CRIMES es el mas repartido en términos de precisión, con todos los distritos en torno al 65%. Esto lo único que indica es que las conclusiones llevadas a cabo anteriormente deben ser tomadas con precaución, y solo asegurar las que son muy seguras, como que los mas afectados son el 1,4,7,18 y 19.

Por ultimo, WEAPONS VIOLATION. Tiene un porcentaje de acierto del 75%, bastante alto. Los distritos mas destacados son el 7, 8 y 10, cosa que difiere un poco de las hipótesis anteriores, ya que hay mas casos en otros mas al sur, y en el 11. Aun así, en líneas generales se ve que los resultados están mas o menos en linea con lo esperado, el crimen está bastante repartido pero contenido en una zona, que es el sur. Esto hace que el modelo consiga clasificar los crímenes de una forma sencilla.



En líneas generales, estos métodos usados a lo largo de este estudio muestran tendencias que confirman la hipótesis inicial que se pueden clasificar los distritos y crímenes basándose en el otro. Bien es cierto que, a pesar de que los primeros métodos sean muy consistentes en cuanto a relaciones y cantidad de eventos, con este último método se ve que puede que no sean tan ciertos. Por este motivo, creo que se deberían tomar en serio las relaciones fuertes, pero las débiles no, ya que eso podría llevar decisiones erróneas.

## 6. Conclusión

El objetivo de este estudio y por tanto de este informe es proporcionar al gobierno de la ciudad y, en caso de pedirlo, a diferentes empresas de seguridad que lo requieren para sus clientes. Por todas las razones expuestas anteriormente voy a mostrar en que distritos se deben emplear prácticas específicas según que crimen, para reducir la criminalidad en vez de generalmente, como se ha hecho en los últimos tiempos. El objetivo de este estudio no es especificar las técnicas que se deberán utilizar para reducir los crímenes, ese será el trabajo de los miembros designados por el gobierno o las empresas, designando mas recursos, usando personal especializado en cada crimen etc.

Por ello, se mostrarán las relaciones mas fuertes e importantes descubiertas en la siguiente tabla.

A continuación de la tabla, se dos cosas, un pequeño resumen de cada crimen y también se comentarán peculiaridades encontradas a lo largo del estudio que no signifiquen relaciones muy fuertes, pero si signifiquen alguna posible modificación en el enfoque que el cliente le quiera dar a su método para reducir la criminalidad.

	<b>Arresto</b>	<b>Distrito</b>	<b>Seguridad</b>
<b>Assault</b>	Mayoritariamente No (Menos el distrito 4)	7 y 4, en menor medida 3,5 y 6.	Media
<b>Criminal Damage</b>	No	8 y la zona sur.	Alta
<b>Deceptive Practice</b>	No	Zona centro (1,18,19) el 24 y el 8.	Alta
<b>Narcotics</b>	Si	10 y 11.	Alta
<b>Robbery</b>	Mayoritariamente No (En los centricos, 1,12,19,18 Si)	Bastante repartido por la ciudad.	Media-Baja
<b>Theft</b>	No	1,18,19 (24 y 16)	Alta
<b>Violent Crime</b>	No (Mayor en los barrios mas centricos)	Mas en el sur de la ciudad, la costa y el centro.	Media-Baja
<b>Weapons Violations</b>	Si	Sur (4,5,6,7,8) y en menor medida Oeste	Alta

En líneas generales, los crímenes marcados con seguridad ALTA, puedo decir con seguridad que, si se centran medidas para reducirlos en los barrios marcados, se acierta con que tipo de crimen se esta trabajando. Es decir, me centraría primero en esos crímenes, luego en ASSAULT y por ultimo, ROBBERY y VIOLENT CRIMES, pero siempre teniendo en cuenta las siguientes recomendaciones.

Aun así, merece la pena comentar cada crimen en particular con sus peculiaridades.

ASSAULT es un crimen no muy común en la ciudad, mayormente repartido por el sur, aunque con bastante foco en los distritos 4 y 7, por lo que puede merecer la pena empezar por ahí. El porcentaje de arresto es muy bajo, a excepción del distrito 4, casualmente el mas afectado, que tiene un porcentaje de arresto mayor.

CRIMINAL DAMAGE es mas común y se centra en el distrito 8 y sus distritos alrededor, con otros focos en el centro de la ciudad y el distrito 24.

DECEPTIVE PRACTICE es también bastante común, y esta casi en su totalidad en el centro de la ciudad, distritos 1,18 y 19. También se encuentran focos en el 24 y el 8.

NARCOTICS es el crimen que mas seguridad muestra en su análisis. Se centra casi en su totalidad en los distritos 10 y 11. Es el crimen con mayor tasa de arrestos, siendo de casi el 100%, por lo que recomendaría en técnicas de concienciación de los vecinos de esos distritos mas que en técnicas para arrestar mas frecuentemente.

ROBBERY es el crimen que mas se reparte por la ciudad. No podría asegurar barrios específicos, pero si indicaría que la tasa de arrestos es mayor en los barrios céntricos, por lo que centraría esfuerzos en otras zonas.

THEFT es, de largo, el crimen mas afectado en la ciudad y además tiene una tasa de arresto bajísima, indicando que la gran mayoría de criminales que llevan a cabo estos no son encontrados. La grandísima mayoría de estos crímenes se centran en el centro de la ciudad, muy probablemente por la grandísima densidad de personas y edificios en la zona y, por la mayor facilidad adquisitiva.

VIOLENT CRIMES es un crimen con una tasa muy baja de arrestos, aunque crece conforme nos acercamos al centro de la ciudad. Se reparte en su mayoría por el sur de la ciudad, incorporando los distritos céntricos 1 y 18.

WEAPONS VIOLATION es el crimen menos común de la ciudad, pero de los mas peligrosos. Sin embargo, tiene una tasa de acierto muy alta. Se reparte en su mayoría por el sur.

Una peculiaridad es la siguiente, hay dos focos de crimen en la ciudad, aparte del obvio del centro, y son el 8 y el 24. Como se ve que el sur de la ciudad es lo mas afectado de la ciudad, el foco en el distrito 8 no me sorprende. Sin embargo, el 24 requiere atención porque significa que hay un foco en el norte, la zona mas segura.

El enfoque que recomendaría es el siguiente, no se va a solucionar el crimen de una de las ciudades con mas criminalidad de Estados Unidos de un día a otro con un informe. Lo primero que haría será centrar esfuerzos en el distrito 24, haciendo del norte entero la zona segura. Una vez hecho eso, recomendaría ir por cada crimen y lanzar campañas y estrategias en distritos con bajo índice de arrestos y para crímenes con bajo crimen de arresto para reducirlos, la razón para ello es lo siguiente, los dos crímenes que menos ocurren, NARCOTICS y WEAPONS VIOLATION son los que mas tasa de arresto tienen, por lo que, si se incrementa la tasa de arresto en los crímenes con una tasa baja, los criminales tendrán mas miedo de cometer el crimen.

En líneas generales, hacer un barrido por crimen desde el sur, centrándose en distritos con bajo índice de arresto, con campañas para incrementar los arrestos (Sin ser necesariamente una mayor presencia policial física en las calles). En crímenes como THEFT o DECEPTIVE PRACTICE, se debería empezar por el centro ya que la grandísima mayoría de casos son en esa zona.

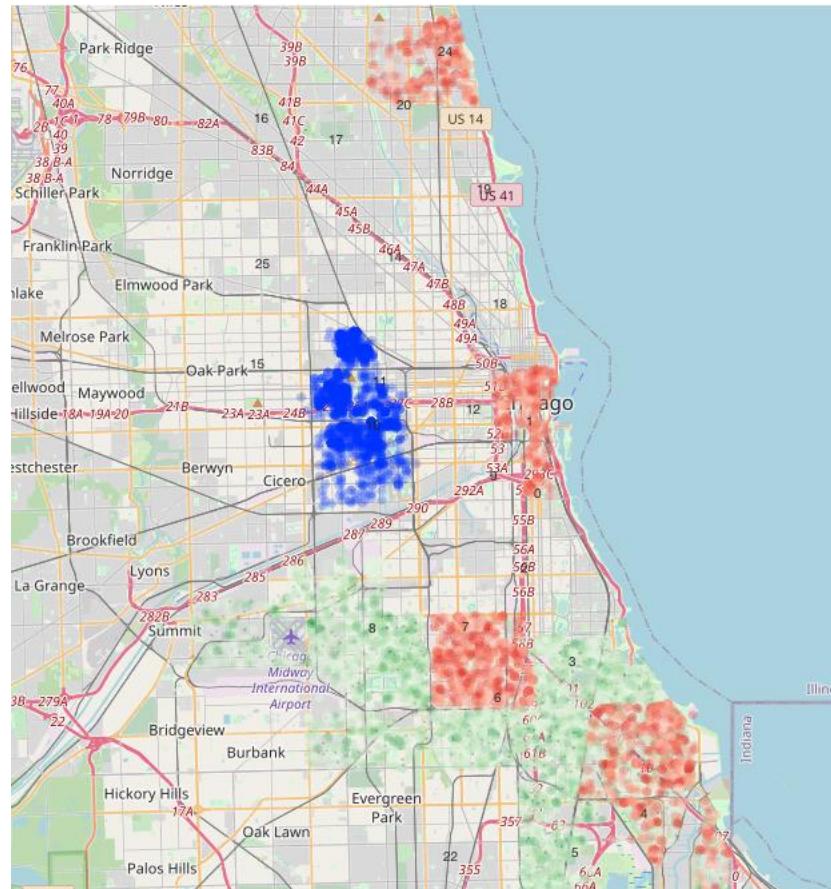
Para los crímenes y distritos con alto porcentaje de arrestos, la campaña que lanzaría sería de concienciación de los vecinos, centrándose en los distritos, enviando los recursos solo a esos distritos y centrados en aquellos crímenes. Por ejemplo, realizaría una campaña de concienciación en el distrito 11 que desarrollara el mensaje de peligrosidad de los narcóticos, y lo mismo en los distritos 3,4,5,6 contra WEAPONS VIOLATION.

Una vez el resto de crímenes tengan una tasa de arrestos alta, combinado con la reducción en casos, realizaría campañas como las mencionadas para los otros distritos.

En los siguientes mapas muestro la distribución recomendada

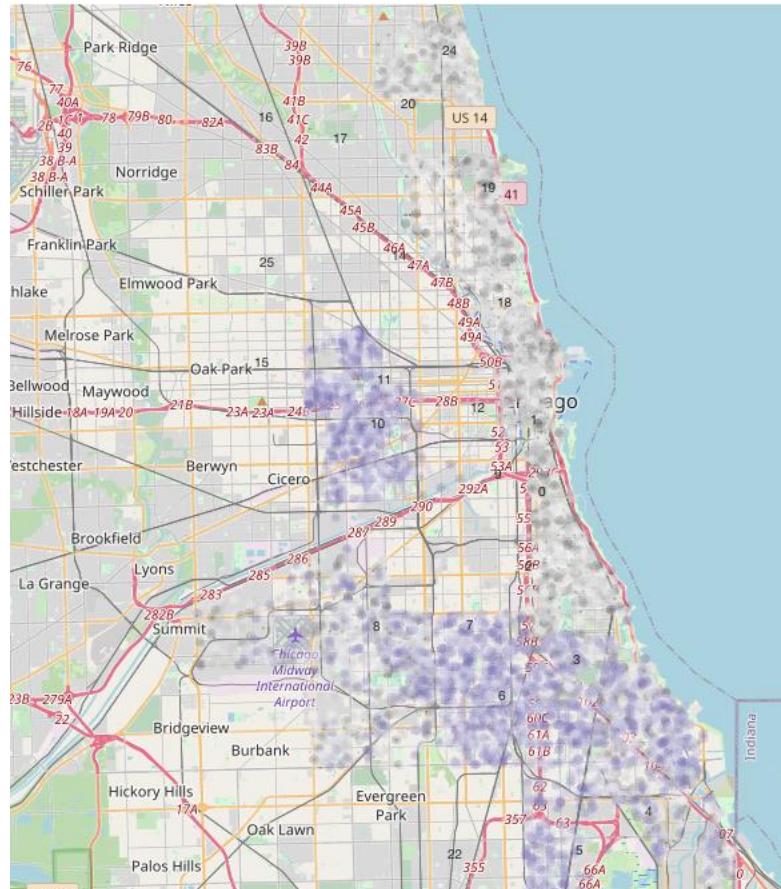
## ASSAULT, NARCOTICS y CRIMINAL DAMAGE

Assault se corresponde con los puntos rojos, Narcotics con los azules y Criminal Damage con los verdes, como se ve, algunos distritos se superponen, decisiones de prioridad se han recomendado anteriormente y se dejan las decisiones finales al cliente.



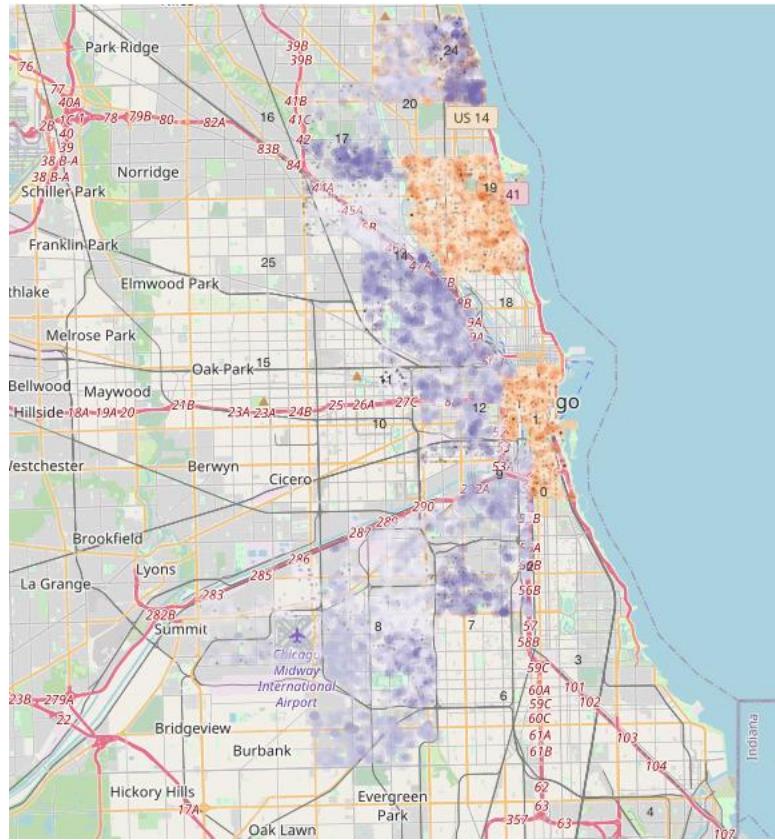
## WEAPONS VIOLATION y VIOLENT CRIMES

Grises se corresponden con Violent Crimes y morados con Weapons Violation. Tambien ocurre superposición.



## **THEFT y DISTRITOS SIN RELACION APARENTE**

Naranjas corresponden a THEFT y morados a los otros distritos, se ve claramente el alto numero en el 24 (Por densidad) y en el 8 (Por tamaño de área).



## DECEPTIVE PRACTICE

