MÁSTER EN DATA SCIENCE,
BIG DATA & BUSINESS ANALYTICS
2023/2024

TAREA DE BASE DE DATOS NoSQL

INSTITUCIÓN: UCM

ESTUDIANTE: Hao, Qi Xu

Índice

INTRODUCCIÓN

El informe tiene como objetivo el análisis de delitos/crímenes que tomaron lugar y fueron denunciados[1] en la ciudad de Los Ángeles, Estados Unidos, con el cual se quiere identificar patrones en estos delitos bajo varias dimensiones, como la temporal, la ubicación geográfica, la tipología de los delitos y el perfil de las víctimas. Dicho análisis puede ayudar a mejorar a comprender la dinámica de los delitos en esta ciudad e identificar áreas de preocupación y contribuir a la formulación de estrategias de prevención.

Para ello, se han recopilado datos sobre delitos denunciados que tomaron lugar desde el 2020 hasta el 1 de noviembre de 2023 (última actualización) proporcionados por el departamento de policía de la ciudad de Los Ángeles: *https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8*

IMPORTACIÓN DE DATOS

La colección de datos cuenta con cerca de 830 mil registros sobre delitos denunciados de los últimos 4 años, desde el 2020 hasta la última actualización de datos: 1 de noviembre de 2023.

Como no es posible descargar el fichero en formato JSON directamente, se ha accedido al API[2] para solicitar los datos requeridos para el estudio. Para evitar posibles problemas de rendimiento relacionados con la solicitud de grandes conjuntos de datos, en vez de una sola consulta, se ha realizado la solicitud en 5 consultas SoQL[3] distintas, pidiendo 166,000 registros en cada una, que posteriormente fueron guardados en 5 ficheros en formato JSON.

Una vez importados los 5 ficheros JSON en MongoDBCompass, observamos que hubo en total 829,778 delitos denunciados que ocurrieron durante los últimos 4 años:

```
db.crimes.find({}).count()
```

```
1  829778
```

[1] Es importante tener en consideración que el número real de delitos cometidos es mayor que los delitos denunciados dado que no siempre llegan a ser denunciados por diversos motivos.

[2] The Socrata Open Data API (SODA). API Endpoint: https://data.lacity.org/resource/2nrs-mtv8.json

[3] **1**: https://data.lacity.org/resource/2nrs-mtv8.json?$limit=166000
2: https://data.lacity.org/resource/2nrs-mtv8.json?$limit=166000&$offset=166000
3: https://data.lacity.org/resource/2nrs-mtv8.json?$limit=166000&$offset=332000
4: https://data.lacity.org/resource/2nrs-mtv8.json?$limit=166000&$offset=498000
5: https://data.lacity.org/resource/2nrs-mtv8.json?$limit=166000&$offset=664000

LOS CAMPOS DE CADA DOCUMENTO (DELITO)

Aquellos campos con fondo gris son los campos de interés que serán utilizados para el análisis posterior. No se va a incluir detalles como la cardinalidad y ejemplo para campos excluidos para análisis posterior.

- **dr_no**: número de expediente oficial compuesto por un año de 2 dígitos, un identificador de área y 5 dígitos.
- **date_rptd**: fecha de la denuncia del delito.
  - Ejemplo: '*2019-04-19T00:00:00.000*'
- **date_occ**: fecha cuando ocurrió el delito.
  - Ejemplo: '*2019-04-19T00:00:00.000*'
- **time_occ**: la hora cuando ocurrió el delito, en formato militar.
  - Ejemplo: '*1900*'[4]
- **area**: código de las 21 Áreas Geográficas o Divisiones Policiales que hace referencia a un punto de referencia o a la comunidad circundante de la cual son responsables.
  - Ejemplo: '*01*'
- **area_name**: nombre de cada una de las 21 Áreas Geográficas.
  - Cardinalidad: 21
  - Ejemplo: 'Central' (hace referencia al código 01 en "area")
- **rpt_dist_no**: Código de cuatro cifras que representa una subzona dentro de un Área Geográfica.
- **part_1_2**: toma como valores 1 o 2. Delitos de tipo 1 (part I) son aquellos delitos contra la persona o contra la propiedad (delitos de homicidio, violación, tráfico de humanos, robo, entre otros). Delitos tipo 2 se refieren a los delitos restantes (vandalismo, prostitución abuso de drogas, otros asaltos, entre otros).[5]
- **crm_cd**: código que indica el delito principal cometido.
- **crm_cd_desc**: descripción del código del delito principal cometido.
  - Cardinalidad: 138
  - Ejemplo: '*BATTERY - SIMPLE ASSAULT*'
- **mocodes**: código de 4 dígitos asociado al Modus Operandi; actividades asociadas con el sospechoso en la comisión del delito y los contextos en los que se produjo. Un registro puede tener varios mocodes.[6]
  - Ejemplo: '0319 0416'
    - 0319: Profanity Used
    - 0416: Hit-Hit w/ weapon
    - 0379: Turns off lights/electricity

[4] El equivalente en el formato estándar: 19:00
[5] Ver Complete_user_manual.pdf para más información.
[6] Las descripciones de los códigos están en MO_CODES_Numerical.pdf.

- **vict_age**: edad de la víctima.
- **vict_sex**: el sexo de la víctima.
    - ❖ Cardinalidad: 5
        - ‘F’ (Female), ‘M’ (Male), ‘X’ (Unknown), ‘H’ , ‘N’[7]
- **vict_descent**: código de una letra que representa cada grupo étnico.
    - ❖ Cardinalidad: 20
        - **A** - Other Asian ,**B** – Black, **C** – Chinese, **D** – Cambodian, **F** - Filipino **G** – Guamanian, **H** - Hispanic/Latin/Mexican, **I** - American Indian/Alaskan Native, **J** – Japanese, **K** – Korean, **L** – Laotian, **O** – Other, **P** - Pacific Islander, **S** – Samoan, **U** – Hawaiian, **V** – Vietnamese, **W** – White, **X** - Unknown, **Z** - Asian Indian
- **premis_cd**: código de 3 dígitos que representa el tipo de estructura, vehículo o ubicación donde ocurrió el delito.
- **premis_desc**: la descripción del tipo de estructura, vehículo o ubicación donde ocurrió el delito.
    - ❖ Cardinalidad: 306
    - ❖ Ejemplo: *‘BANK’*
- **weapon_used_cd**: código de 3 dígitos que representa el tipo de arma utilizada en el delito.
- **weapon_desc**: descripción del tipo de arma utilizada en el delito.
    - ❖ Cardinalidad**:** 79
    - ❖ Ejemplo: *‘RIFLE’*
- **status**: abreviatura del estado en el que se encuentra el caso de delito.
- **status_desc**: descripción del estado en el que se encuentra el delito.
    - ❖ Cardinalidad: 6
        - ‘*Invest Cont’*: investigación en curso, recopilación pruebas, no hay arresto.
        - ‘*Adult Arrest’*, ‘*Juv Arrest’*: adulto/joven arrestado en relación con el caso.
        - ‘*Adult Other’*, ‘*Juv Other’*: otros estados que no implican arresto.
        - ‘*UNK*’: estado desconocido / no especificado.
- **location**: dirección de la calle aproximado donde ocurrió el delito.
- **lat**: coordenada de latitud donde ocurrió el delito.
- **lon**: coordenada de longitud donde ocurrió el delito.

Se han elegido todos los campos con la descripción de los valores en vez de sus versiones codificados en dígitos.

Se escoge el campo “area_name” como campo principal geográfico por descarte. Por un lado, “location” presenta una cardinalidad muy alta, llegando a tener 63,821 valores únicos, mientras que “area_name” solo presenta 21 categorías. Por otro lado, los campos asociados a las coordenadas

[7] No se especifica en ningún lado de la documentación lo que H y N representan.

geográficas de los delitos no se van a poder aprovechar sin poder visualizarlas gráficamente en MongoDB.

Finalmente, se quiere comprobar si cada documento representa un delito distinto asociado a una sola víctima o un mismo documento (delito) puede estar asociado a varias víctimas; es decir, si la colección presenta duplicados del mismo delito con distintas víctimas. Se procede a comprobar con el conteo por cada "dr_no", el número de expediente oficial de cada delito denunciado:

```
db.crimes.aggregate([
  {$group: {
    _id: "$dr_no",
    count: {$sum: 1}
  }},
  {$match: {count: {$gt: 1}}}
]).count()
```

1	3466

Al parecer, hay 3,466 registros duplicados por el campo "dr_no".

```
db.crimes.find({dr_no: "201208887"})
```

	_id	dr_no	date_rptd	date_occ	time_occ	area	area_name	rpt_dist_no	part_1_2	crm_cd
1	6547ec76328ee544184	201208887	2020-03-10T00:00:00.000	2020-03-09T00:00:00.000	0030	12	77th Street	1213	1	341
2	6547ec81328ee544184	201208887	2020-03-10T00:00:00.000	2020-03-09T00:00:00.000	0030	12	77th Street	1213	1	341

	_id	crm_cd_desc	mocodes	vict_age	vict_sex	vict_descent	premis_cd	premis_desc	status	status_desc	crm_cd
1	6547ec76328ee544184	THEFT-GRAND ($950.01 &	1602 0344 1822 0325	0	X	X	708	CHURCH/CHAPEL (CHAN	IC	Invest Cont	341
2	6547ec81328ee544184	THEFT-GRAND ($950.01 &	1602 0344 1822 0325	0	X	X	708	CHURCH/CHAPEL (CHAN	IC	Invest Cont	341

Después de haber revisado un par de aquellos documentos duplicados al azar, se observa que todos ellos contienen el mismo valor para todos los campos disponibles, incluso en las características de la víctima asociada a cada crimen. Esto sugiere que cada documento (delito) es único y está asociada a una sola víctima. Como los delitos con el mismo número de expediente no aportan ninguna información extra, se procede a eliminar dichos documentos duplicados[8]

```
const bulkOps = []
db.crimes.aggregate([
  {
    $group: {
      _id: "$dr_no",
      uniqueIds: { $addToSet: "$_id" },
      count: { $sum: 1 }
    }
  },
  {
    $match: {
      count: { $gt: 1 }
```

[8] La query para eliminar duplicados fue posible gracias en gran medida a la ayuda de CHATGPT.

```
    }
  }
]).forEach(function (doc) {
  doc.uniqueIds.shift()
  bulkOps.push({
    deleteOne: {
      filter: { _id: { $in: doc.uniqueIds } }
    }
  })
})
if (bulkOps.length > 0) {
  db.crimes.bulkWrite(bulkOps)
};
```


```
db.crimes_modificado.count()
```

Una vez eliminado los registros duplicados, contamos con un total de 826,312 documentos.

MODIFICACIÓN Y CREACIÓN DE UNA NUEVA COLECCIÓN

Nótese que algunos campos podrían ser modelizados de manera más adecuada y que todos los campos son de tipo string:

db.crimes.find({}).limit(1)


```
{
    "_id" : ObjectId("6547ec5e328ee54418463f00"),
    "dr_no" : "010304468",
    "date_rptd" : "2020-01-08T00:00:00.000",
    "date_occ" : "2020-01-08T00:00:00.000",
    "time_occ" : "2230",
    "area" : "03",
    "area_name" : "Southwest",
    "rpt_dist_no" : "0377",
    "part_1_2" : "2",
    "crm_cd" : "624",
    "crm_cd_desc" : "BATTERY - SIMPLE ASSAULT",
    "mocodes" : "0444 0913",
    "vict_age" : "36",
    "vict_sex" : "F",
    "vict_descent" : "B",
    "premis_cd" : "501",
    "premis_desc" : "SINGLE FAMILY DWELLING",
    "weapon_used_cd" : "400",
    "weapon_desc" : "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)",
    "status" : "AO",
    "status_desc" : "Adult Other",
    "crm_cd_1" : "624",
    "location" : "1100 W  39TH                          PL",
    "lat" : "34.0141",
    "lon" : "-118.2978"
}
```


Por ello, se realizarán los siguientes cambios en los campos de interés:

- Los campos relacionados a las características de las víctimas pasan a formar un nuevo campo "**victim**" con los campos de sus características embedidos. Además, se le cambia también el tipo del campo "vict_age" previamente a integer.
- Se crea un nuevo campo "**MOcodes**" a partir del campo "mocodes" que contiene en su array los códigos de 4 dígitos que representan los métodos y las actividades que han utilizado el criminal en el delito, así como los contextos en los que se produjo.
- Se crea un nuevo campo "**time**" a partir de "time_occ" en el cual se convierte la hora en formato militar al formato estándar; por ejemplo, de "2230" a "22:30".
- Una vez creada "time", se le concatena con el campo "date_occ" para formar un nuevo campo "**date_occurrence**", el cual contiene la fecha y la hora que tuvo lugar cada delito.
- Por último, creamos un nuevo campo "**date_reported**" a partir de "date_rptd" que tiene datetime como tipo de datos.

Cabe destacar que todas las modificaciones y nuevos campos creados se han realizado en un único pipeline de agregación. Además, se ha optado por crear una nueva colección llamada "**crimes_modificado**" con las modificaciones mencionadas aplicadas y filtradas solamente por los campos de interés, listo para la fase de análisis:

```
db.crimes.aggregate([
  {$addFields: {
    "victim.age": {$toInt: "$vict_age"},
    "victim.sex": "$vict_sex",
    "victim.descent": "$vict_descent",
    "MOcodes": {$split: ["$mocodes", " "]},
    "time": {$concat: [{$substr: ["$time_occ", 0, 2]}, ":", {$substr: ["$time_occ", 2, 2]}]},
    "date_reported": {$toDate: "$date_rptd"},
    "date_occ": {$toDate: "$date_occ"}
  }},
  {$addFields: {
    "date_occurrence": {$toDate: {$concat: [{$substr: ["$date_occ", 0, 10]}, "T", "$time", ":00"]}}
  }},
  {$project: {
    "_id": false,
    "date_occurrence": true,
    "date_occ": true,
    "date_reported": true,
    "area_name": true,
    "status_desc": true,
    "crm_cd_desc": true,
    "part_1_2": true,
    "MOcodes": true,
    "weapon_desc": true,
    "premis_desc": true,
    "victim.age": true,
    "victim.sex": true,
    "victim.descent": true
  }},
  {$out: "crimes_modificado"}
])
```

db.crimes_modificado.find({}).limit(1)


```
{
    "_id" : ObjectId("6547f076d9eee3f6725f002f"),
    "date_occ" : ISODate("2020-01-08T01:00:00.000+01:00"),
    "area_name" : "Southwest",
    "part_1_2" : "2",
    "crm_cd_desc" : "BATTERY - SIMPLE ASSAULT",
    "premis_desc" : "SINGLE FAMILY DWELLING",
    "weapon_desc" : "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)",
    "status_desc" : "Adult Other",
    "victim" : {
        "age" : 36,
        "sex" : "F",
        "descent" : "B"
    },
    "MOcodes" : [ "0444", "0913" ],
    "date_reported" : ISODate("2020-01-08T01:00:00.000+01:00"),
    "date_occurrence" : ISODate("2020-01-08T23:30:00.000+01:00")
}
```


FASE DE ANÁLISIS

El estudio de los crímenes en LA (Los Ángeles) se realizará por sus diferentes dimensiones mencionadas anteriormente:

Dimensión Temporal

- En primer lugar, queremos conocer cuántos delitos ocurrieron desde 2020 hasta la fecha de la última actualización, 1 de noviembre del 2023:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: {
      año: {$year: "$date_occurrence"}
    },
    count: {$sum: 1}
  }},
  {$sort: {_id: 1}}
])
```

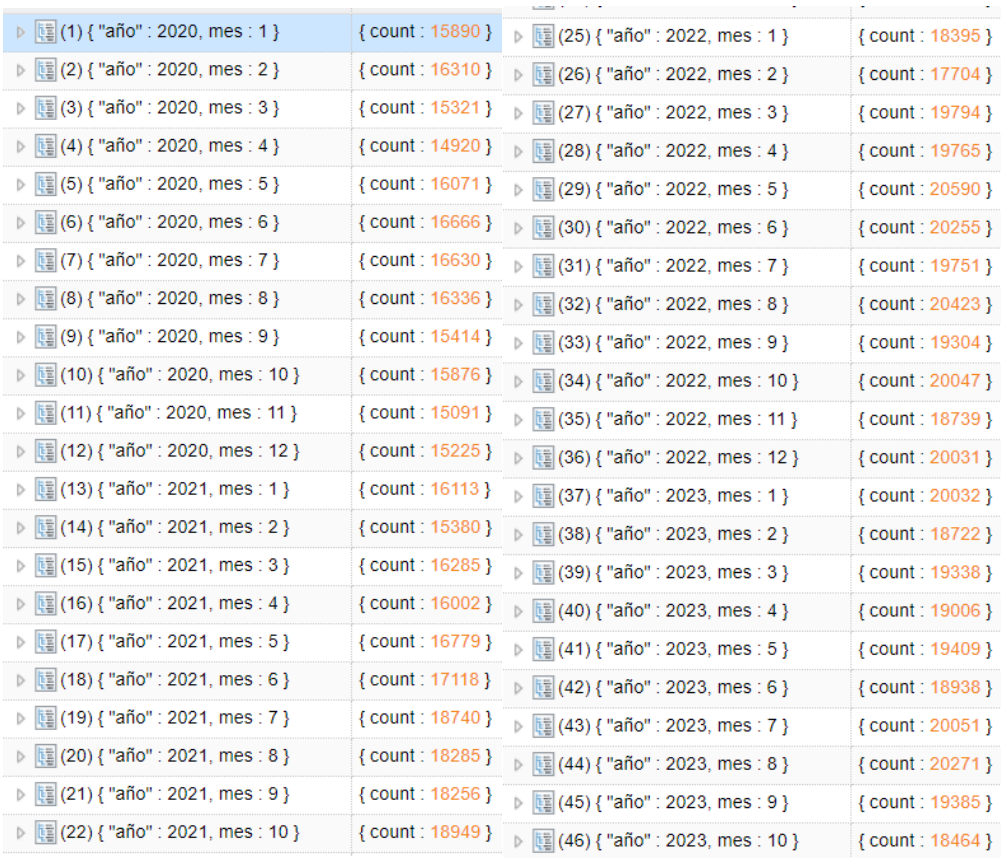
Key	Value
(1) { "año" : 2020 }	{ count : 195961 }
(2) { "año" : 2021 }	{ count : 209379 }
(3) { "año" : 2022 }	{ count : 234289 }
(4) { "año" : 2023 }	{ count : 186683 }

Se contempla que la incidencia absoluta de delitos aumenta gradualmente cada año desde el 2020 hasta al 2022, incluidos. Hay cierta incertidumbre en si la incidencia de delitos en el año 2023 va a ser mayor o menor que el año anterior, ya que no se han registrado todavía los datos de los meses de noviembre y diciembre. No obstante, la ausencia de los datos del tamaño de la población de la ciudad de Los Ángeles (LA) en esos años para ajustar la incidencia de crímenes no nos permite

obtener conclusiones válidas. La incidencia criminalidad absoluta, a diferencia de la tasa de criminalidad per cápita, no tiene en cuenta la relación que existe entre el nivel de la población y la tasa de criminalidad. El aumento de la tasa de criminalidad en esos años podría ser atribuido a un aumento significativo de la población en esos períodos.

- ¿La incidencia de crímenes tiende a concentrarse en algunos meses del año?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: {
      año: {$year: "$date_occurrence"},
      mes: {$month: "$date_occurrence"}
    },
    count: {$sum: 1}
  }},
  {$sort: {_id: 1}}
])
```



(1) { "año" : 2020, mes : 1 }	{ count : 15890 }	(25) { "año" : 2022, mes : 1 }	{ count : 18395 }
(2) { "año" : 2020, mes : 2 }	{ count : 16310 }	(26) { "año" : 2022, mes : 2 }	{ count : 17704 }
(3) { "año" : 2020, mes : 3 }	{ count : 15321 }	(27) { "año" : 2022, mes : 3 }	{ count : 19794 }
(4) { "año" : 2020, mes : 4 }	{ count : 14920 }	(28) { "año" : 2022, mes : 4 }	{ count : 19765 }
(5) { "año" : 2020, mes : 5 }	{ count : 16071 }	(29) { "año" : 2022, mes : 5 }	{ count : 20590 }
(6) { "año" : 2020, mes : 6 }	{ count : 16666 }	(30) { "año" : 2022, mes : 6 }	{ count : 20255 }
(7) { "año" : 2020, mes : 7 }	{ count : 16630 }	(31) { "año" : 2022, mes : 7 }	{ count : 19751 }
(8) { "año" : 2020, mes : 8 }	{ count : 16336 }	(32) { "año" : 2022, mes : 8 }	{ count : 20423 }
(9) { "año" : 2020, mes : 9 }	{ count : 15414 }	(33) { "año" : 2022, mes : 9 }	{ count : 19304 }
(10) { "año" : 2020, mes : 10 }	{ count : 15876 }	(34) { "año" : 2022, mes : 10 }	{ count : 20047 }
(11) { "año" : 2020, mes : 11 }	{ count : 15091 }	(35) { "año" : 2022, mes : 11 }	{ count : 18739 }
(12) { "año" : 2020, mes : 12 }	{ count : 15225 }	(36) { "año" : 2022, mes : 12 }	{ count : 20031 }
(13) { "año" : 2021, mes : 1 }	{ count : 16113 }	(37) { "año" : 2023, mes : 1 }	{ count : 20032 }
(14) { "año" : 2021, mes : 2 }	{ count : 15380 }	(38) { "año" : 2023, mes : 2 }	{ count : 18722 }
(15) { "año" : 2021, mes : 3 }	{ count : 16285 }	(39) { "año" : 2023, mes : 3 }	{ count : 19338 }
(16) { "año" : 2021, mes : 4 }	{ count : 16002 }	(40) { "año" : 2023, mes : 4 }	{ count : 19006 }
(17) { "año" : 2021, mes : 5 }	{ count : 16779 }	(41) { "año" : 2023, mes : 5 }	{ count : 19409 }
(18) { "año" : 2021, mes : 6 }	{ count : 17118 }	(42) { "año" : 2023, mes : 6 }	{ count : 18938 }
(19) { "año" : 2021, mes : 7 }	{ count : 18740 }	(43) { "año" : 2023, mes : 7 }	{ count : 20051 }
(20) { "año" : 2021, mes : 8 }	{ count : 18285 }	(44) { "año" : 2023, mes : 8 }	{ count : 20271 }
(21) { "año" : 2021, mes : 9 }	{ count : 18256 }	(45) { "año" : 2023, mes : 9 }	{ count : 19385 }
(22) { "año" : 2021, mes : 10 }	{ count : 18949 }	(46) { "año" : 2023, mes : 10 }	{ count : 18464 }

No parece ser el caso. No parece haber una diferencia significativa en la incidencia de delitos en algún mes en particular del año en los últimos 4 años.

- ¿Es más probable que ocurra un delito en una franja horaria específica a lo largo del día?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: {
      hora: {$hour: "$date_occurrence"}
    },
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {_id: 1}}
])
```

(1) { hora : 0 }	{ count : 34123, percent : 4.1295539699290345 }
(2) { hora : 1 }	{ count : 25061, percent : 3.032873781331991 }
(3) { hora : 2 }	{ count : 21198, percent : 2.565374822101095 }
(4) { hora : 3 }	{ count : 18444, percent : 2.232086669442051 }
(5) { hora : 4 }	{ count : 15649, percent : 1.8938367105887366 }
(6) { hora : 5 }	{ count : 14402, percent : 1.74292519048495 }
(7) { hora : 6 }	{ count : 19004, percent : 2.2998576808759887 }
(8) { hora : 7 }	{ count : 21365, percent : 2.5855851058680015 }
(9) { hora : 8 }	{ count : 30533, percent : 3.695093378772183 }
(10) { hora : 9 }	{ count : 29817, percent : 3.6084432998673623 }
(11) { hora : 10 }	{ count : 35398, percent : 4.283854040604517 }
(12) { hora : 11 }	{ count : 35867, percent : 4.34061226268044 }
(13) { hora : 12 }	{ count : 56003, percent : 6.7774642023836025 }

(13) { hora : 12 }	{ count : 56003, percent : 6.7774642023836025 }
(14) { hora : 13 }	{ count : 37324, percent : 4.5169379120719535 }
(15) { hora : 14 }	{ count : 40276, percent : 4.874187958059426 }
(16) { hora : 15 }	{ count : 43050, percent : 5.209896503983967 }
(17) { hora : 16 }	{ count : 43437, percent : 5.256731113671349 }
(18) { hora : 17 }	{ count : 47940, percent : 5.801682657398174 }
(19) { hora : 18 }	{ count : 48959, percent : 5.925001694275286 }
(20) { hora : 19 }	{ count : 45291, percent : 5.481101569382994 }
(21) { hora : 20 }	{ count : 45899, percent : 5.554681524654126 }
(22) { hora : 21 }	{ count : 42019, percent : 5.085125231147557 }
(23) { hora : 22 }	{ count : 40514, percent : 4.902990637918849 }
(24) { hora : 23 }	{ count : 34739, percent : 4.204102082506365 }

  - Mañana (6-12): 27.60%
  - Tarde (13-18): 31.59%
  - Noche (19-0): 29.27%
  - Madrugada (1-5): 11.46%

Aparte de la madrugada, período el cual hay menos interacción social, no hay una diferencia significativa a primera vista en la incidencia de delitos entre las distintas franjas horarias del día.

- Dado la disponibilidad de la fecha de la denuncia de los delitos se quiere aprovechar para identificar el promedio de días que uno (víctima o testigo) tarda en denunciar el delito después de que dicho evento haya ocurrido:

```
db.crimes_modificado.aggregate([
  {$addFields: {
    rpt_delay_days: {$divide: [{$subtract: ["$date_reported", "$date_occ"[9]]}, 86400000[10]]}
```

[9] Se utiliza “date_occ” en vez de “date_occurrence” en la resta porque tanto el campo “date_rptd” como “date_occ” no contienen la información de la hora.
[10] La resta de dos fechas en Mongo es devuelta en milisegundos y un día está compuesto de 8,6400,000 milisegundos.

```
    }},
    {$group: {
      _id: null,
      avg_delay_days: {$avg: "$rpt_delay_days"},
      median_delay_days: {$median: {
       input: "$rpt_delay_days",
       method: 'approximate'
      }},
      min_delay_days: {$min: "$rpt_delay_days"},
      max_delay_days: {$max: "$rpt_delay_days"}
    }}
])
```

	_id	avg_delay_days	median_delay_days	min_delay_days	max_delay_days
1	null	10.14161963035754	1.0	0.0	1379.0

Se intuye la presencia de valores extremos en el retraso de la denuncia de los delitos, puesto que su promedio difiere significativamente de su mediana. Podemos confirmarlo al observar el retraso mínimo y máximo, de 0 a 1,379 días (casi cuatro años), respectivamente. Por ello, la medida de centralidad más adecuada para la variable "rpt_delay_days" es la mediana, ya que no es afectada por los valores extremos. Las víctimas o testigos de los delitos denuncian los crímenes a lo largo de un día de mediana, después de su incidencia.

- Por último, se quiere identificar cuáles son los principales delitos cometidos con al menos 100 observaciones que tuvieron de mediana mayores días de retraso en su denuncia:

```
db.crimes_modificado.aggregate([
    {$addFields: {
      rpt_delay_days: {$divide: [{$subtract: ["$date_reported", "$date_occ"]}, 86400000]}
    }},
    {$group: {
      _id: {
         delito_principal: "$crm_cd_desc"
      },
      count: {$sum: 1},
      avg_delay_days: {$avg: "$rpt_delay_days"},
      median_delay_days: {$median: {
       input: "$rpt_delay_days",
       method: 'approximate'
      }},
      min_delay_days: {$min: "$rpt_delay_days"},
      max_delay_days: {$max: "$rpt_delay_days"}
```

```
    }},
    {$match: {count: {$gte: 100}}},
    {$sort: {median_delay_days: -1}},
    {$limit: 10}
])
```

	_id	count	avg_delay_days	median_delay_days	min_delay_days	max_delay_days
1	{ delito_principal : "SEX OFFENDER REGISTRANT OUT OF COMPLIANCE" }	837	109.28793309438471	48.0	0.0	1160.0
2	{ delito_principal : "SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ FRGN OBJ" }	900	108.39222222222222	38.0	0.0	1170.0
3	{ delito_principal : "EMBEZZLEMENT, GRAND THEFT ($950.01 & OVER)" }	3137	56.34077143767931	32.0	0.0	1217.0
4	{ delito_principal : "EMBEZZLEMENT, PETTY THEFT ($950 & UNDER)" }	117	60.94017094017094	19.0	0.0	1288.0
5	{ delito_principal : "DOCUMENT FORGERY / STOLEN FELONY" }	2872	55.06267409470752	17.0	0.0	1050.0
6	{ delito_principal : "CREDIT CARDS, FRAUD USE ($950.01 & OVER)" }	103	46.66019417475728	11.0	0.0	569.0
7	{ delito_principal : "CRM AGNST CHLD (13 OR UNDER) (14-15 & SUSP 10 YRS OLDER)" }	1507	114.19309887193095	9.0	0.0	1337.0
8	{ delito_principal : "UNAUTHORIZED COMPUTER ACCESS" }	479	49.805845511482254	6.0	0.0	1330.0
9	{ delito_principal : "THEFT OF IDENTITY" }	52122	51.82308813936533	6.0	0.0	1379.0
10	{ delito_principal : "BUNCO, GRAND THEFT" }	5354	31.603847590586476	4.0	0.0	1213.0

La consulta señala que los crímenes que presentan mayores días de retraso en la denuncia fueron principalmente aquellos relacionados con abusos sexuales (1 y 2) y malversación de fondos (3 y 4), así como hurtos mayores de $950. Los tres primeros delitos principales fueron denunciados de mediana más de un mes después de que ocurrieran. En el caso de los abusos sexuales puede deberse a la vergüenza y al estigma asociado con ser víctimas de tales crímenes, y/o la dependencia económica o emocional de las víctimas del agresor. En cuanto a las malversaciones y hurtos mayores, la tardía denuncia podría ser causado por la complejidad de detección y la recopilación de suficientes pruebas de dichos delitos, especialmente en el caso de las malversaciones puesto que suelen ser personas de altos cargos dentro de las instituciones.

Dimensión Geográfica

- ¿Cuáles son las áreas de la ciudad que concentraron mayor incidencia crímenes):

```
db.crimes_modificado.aggregate([
    {$group: {
        _id: "$area_name",
        count: {$sum: 1}
    }},
    {$project: {
        _id: true,
        count: true,
        percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
    }},
    {$sort: {percent: -1}}
])
```

	_id	count	percent
1	Central	55862	6.760400
2	77th Street	52111	6.306455
3	Pacific	48258	5.840166
4	Southwest	46358	5.610229
5	Hollywood	43810	5.301871
6	Southeast	41948	5.076532
7	Olympic	41745	5.051965
8	Newton	41338	5.002710
9	N Hollywood	41065	4.969672
10	Wilshire	39361	4.763454
11	Rampart	38698	4.683218
12	West LA	37963	4.594269
13	Northeast	35670	4.316771
14	Van Nuys	34962	4.231089
15	West Valley	34573	4.184012
16	Harbor	34209	4.139961
17	Topanga	33594	4.065534
18	Devonshire	33413	4.043630
19	Mission	32863	3.977069
20	Hollenbeck	31014	3.753303
21	Foothill	27497	3.327677

Se aprecia una cierta patrón en el resultado de la consulta. Las áreas de las divisiones policiales con más incidencias absolutas de delitos también son las zonas con mayor densidad de población en la ciudad de Los Ángeles: Central, 77th Street, Pacific, Southwest, Hollywood, entre otros. Similarmente, Áreas como Devonshire o Foothill, con menores incidencias de delitos, están bastante alejadas del centro de la ciudad y tienen menor densidad de población.

Especial atención recibe la zona Central de LA ya que en ella se encuentra "Skid Row", un barrio en el que reside la mayor concentración de gente sin techo del estado de California (y de todo Estados Unidos)[11]. Parte de los delitos en esa area tiene lugar entre la comunidad de gente sin techo en las calles de dicho barrio.

- Se quiere comprobar la hipótesis de que una fracción significativa de los crímenes en LA Central tiene lugar entre la gente sin techo. Esta consulta es posible gracias a que la variable "MOcodes", la cual contiene una array de códigos que representan los contextos bajo los cuales cada delito tuvo lugar, contiene las siguientes detalles:
  - "*1218*": Victim was homeless
  - "*2004*": Suspect is homeless

```
db.crimes_modificado.find({MOcodes: {$all:["1218", "2004"]}}).count()
```

```
1  8352
```

```
db.crimes_modificado.find({area_name: "Central"}).count()
```

```
1  55862
```

```
db.crimes_modificado.aggregate([
  {$match: {MOcodes: {$all:["1218", "2004"]}}},
  {$group: {
    _id: "$area_name",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
```

[11] Skid Row: el infierno de la mayor concentración de indigentes de EE.UU. - BBC News Mundo

```
    percent: {$multiply:[{$divide: ["$count", 8352]}, 100]}
  }},
  {$sort: {count: -1}}
])
```

	_id	count	percent
1	Central	2887	34.566570881226056
2	Pacific	822	9.841954022988505
3	Newton	490	5.866858237547893
4	N Hollywood	487	5.8309386973180075
5	Hollywood	466	5.579501915708812
6	Rampart	403	4.825191570881226
7	Olympic	284	3.400383141762452
8	Harbor	280	3.3524904214559386
9	Devonshire	240	2.8735632183908044
10	West LA	223	2.6700191570881224
11	Van Nuys	219	2.6221264367816093
12	Northeast	214	2.562605363984676
13	77th Street	187	2.238984674329502
14	Hollenbeck	167	1.9995210727969348
15	Wilshire	167	1.9995210727969348
16	Southwest	165	1.9755747126436782
17	West Valley	153	1.8318965517241377
18	Southeast	150	1.7959770114942528
19	Foothill	131	1.5684865900383143
20	Topanga	115	1.3769157088122606
21	Mission	102	1.221264367816092

Hubo un total de 8,352 delitos en los que tanto las víctimas como los sospechosos fueron gente sin techo. Del total de delitos cometidos entre gente sin techo, más de un tercio de ellos, 34.57%, se concentran en LA Central, confirmando la relativa mayor influencia de los sin techos en los crímenes en dicha área. No obstante, estos delitos constituyen solamente el (2,887 / 55,862) x 100 = 5.17% del total de crímenes cometidos en LA Central en estos últimos 4 años.

- ¿En qué espacio/lugar en concreto tuvieron lugar gran parte de los crímenes?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: "$premis_desc",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {percent: -1}},
  {$limit: 10}
])
```

	_id	count	percent
1	STREET	208344	25.213720
2	SINGLE FAMILY DWELLIN	140253	16.973370
3	MULTI-UNIT DWELLING (A	101664	12.303343
4	PARKING LOT	57668	6.9789619
5	OTHER BUSINESS	38920	4.7100852
6	SIDEWALK	35703	4.3207650
7	VEHICLE, PASSENGER/TI	24582	2.9749053
8	GARAGE/CARPORT	16205	1.9611236
9	DRIVEWAY	13436	1.6260201
10	RESTAURANT/FAST FOOI	10579	1.2802670

Se observa que más de la mitad de los crímenes, 54.48%, tienen lugar en 3 lugares específicos: en la calle/vías públicas (25.21%), en viviendas unifamiliares (16.97%), y edificios de viviendas múltiples (12.3%).

Tipología del delito

Los delitos cometidos son en Estados Unidos son clasificados por dos categorías: delitos de Parte I (Part I Offenses) y delitos de Parte II (Part II Offenses). Los delitos de Parte I son considerados más graves: homicidio criminal, violación, asalto agravado, trata de personas, allanamiento de morada, robo de automóviles, entre otros. Por otra parte, los delitos de Parte II son el resto considerados menos graves que los de Parte I: asalto no agravado, vandalismo, prostitución, malversación, abuso de drogas, entre otros.

- ¿Cómo se distribuyen los delitos por esta tipología?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: "$part_1_2",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {percent: -1}}
])
```

	_id	count	percent
1	1	483717	58.5392684
2	2	342595	41.4607315

Según el resultado, hubo más delitos de Parte I que de Parte II en los últimos 4 años, un 17.08% más en concreto.

- ¿Cuáles son los delitos cometidos con mayor incidencia en ambos tipos?:

```
db.crimes_modificado.aggregate([
  {$match: {part_1_2: "1"}},
  {$group: {
    _id: "$crm_cd_desc",
    count: {$sum: 1}
  }},
  {$sort: {count: -1}},
  {$limit: 10}
])
```

	_id	count
1	VEHICLE - STOLEN	87106
2	BURGLARY FROM VEHICLE	50916
3	BURGLARY	50506
4	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	47500
5	THEFT PLAIN - PETTY ($950 & UNDER)	42144
6	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	31744
7	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)	29969
8	ROBBERY	28336
9	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD	27026
10	SHOPLIFTING - PETTY THEFT ($950 & UNDER)	19574

El resultado indica que la mayoría de los delitos de Parte I están relacionados con delitos de propiedad. Muchos de ellos están asociados a robos de vehículos en sí y sus motores.

```
db.crimes_modificado.aggregate([
  {$match: {part_1_2: "2"}},
  {$group: {
    _id: "$crm_cd_desc",
    count: {$sum: 1}
  }},
  {$sort: {count: -1}},
  {$limit: 10}
])
```

	_id	count
1	BATTERY - SIMPLE ASSAULT	66121
2	THEFT OF IDENTITY	52122
3	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	50536
4	INTIMATE PARTNER - SIMPLE ASSAULT	41737
5	VANDALISM - MISDEAMEANOR ($399 OR UNDER)	22148
6	CRIMINAL THREATS - NO WEAPON DISPLAYED	17173
7	TRESPASSING	11738
8	VIOLATION OF RESTRAINING ORDER	10527
9	LETTERS, LEWD - TELEPHONE CALLS, LEWD	6840
10	OTHER MISCELLANEOUS CRIME	5984

A diferencia de los de Parte I, los delitos de Parte II presentan más variedad. Los más comunes son las agresiones físicas, las amenazas de causar daño físico, robo de identidad y vandalismo.

- Finalmente, se quiere conocer las actividades y los contextos más comunes bajos los cuales ocurrieron los delitos durante los últimos 4 años:

```
db.crimes_modificado.aggregate([
  {$unwind: "$MOcodes"},
  {$group: {
    _id: "$MOcodes",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {percent: -1}},
  {$limit: 10}
])
```

	_id	count	percent
1	1822	284031	34.37333597962997
2	0344	243906	29.51742199072505
3	0913	134352	16.259233800307875
4	0329	109814	13.28965330286865
5	0416	107358	12.992429009865521
6	1300	81022	9.80525515785805
7	2000	66020	7.989718169408165
8	0400	65655	7.945545992312832
9	1402	51497	6.232149599666954
10	2004	45759	5.537738771795642

- *1822 (*Stranger[12]*)*: hubo una persona desconocida involucrada en el 34.37% de los delitos.
- *0344 (*Removes vict property*):* hubo un robo/hurto de la propiedad de la víctima en el 29.52% de los casos.
- *0913 (*Victim knew suspect*):* la víctima conocía a su agresor/sospechoso en el 16.26% de los casos.
- *0329 (*Vandalized*)*: hubo destrucción de propiedad en el 13.29% de los casos.
- *0416 (*Hit-Hit w/weapon*)*: la víctima fue agredida con el uso de cualquier arma en el 13% de los casos.
- *1300 (*Vehicle involved*)*: algún vehículo estuvo involucrado en el 9.8% de los casos.
- *2000 (*Domestic violence*)*: hubo violencia doméstica en el 8% de los casos.
- *0400 (*Force used*)*: la fuerza policial fue utilizado en el 7.94% de los casos.
- *1402 (*Evidence booked*)*: se han registrado pruebas/evidencias en relación con el delito en el 6.23% de los casos.
- 2004 (Suspect is homeless): el sospechoso era una persona sin hogar en el 5.54% de los casos.

Perfil de las víctimas

Por último, se quiere identificar el perfil de víctimas más comunes en los delitos cometidos en los últimos 4 años en LA:

- ¿Cuál es la edad media de las víctimas?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: null,
    avg_age: {$avg: "$victim.age"},
    median_age: {$median: {
     input: "$victim.age",
     method: 'approximate'
    }},
    min_age: {$min: "$victim.age"},
    max_age: {$max: "$victim.age"}
  }}
])
```

	_id	avg_age	median_age	min_age	max_age
1	null	29.862662045329124	31.0	-3	120

Se percibe un resultado extraño en la edad mínima de las víctimas; -3. No se conoce si se trata de un problema relacionado con la calidad de datos o un código con un significado específico. Por

[12] No se especifica lo que realmente representa. Intuyo que se refiere a que hubo personas desconocidas involucradas en el crimen.

ello, se utilizará la mediana para evitar la influencia de datos atípicos. Por consiguiente, las víctimas de los delitos tienen de mediana 31 años.

- ¿Cómo es la distribución de los delitos respecto al sexo de la víctima?:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: "$victim.sex",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {percent: -1}}
])
```

	_id	count	percent
1	M	341789	41.363189691061
2	F	304850	36.892844349349886
3	null	107076	12.958301464822004
4	X	72506	8.774651705409095
5	H	90	0.01089176969474
6	-	1	0.00012101966327488891

No se documenta lo que los valores “H” y “-“ representan. Consecuentemente, estos valores no serán interpretados. Se observa que aproximadamente el 41.36% de las víctimas son hombres (M) y el 36.9%, mujeres (F). Por lo tanto, hubo un 4.46% o 36,939 más víctimas hombres que mujeres.

- Por último, se quiere conocer los grupos étnicos más vulnerables en relación con los crímenes:

```
db.crimes_modificado.aggregate([
  {$group: {
    _id: "$victim.descent",
    count: {$sum: 1}
  }},
  {$project: {
    _id: true,
    count: true,
    percent: {$multiply:[{$divide: ["$count", 826312]}, 100]}
  }},
  {$sort: {percent: -1}},
```

```
    {$limit: 10}
])
```

	_id	count	percent
1	H	254050	30.745045454985526
2	W	168723	20.41880064672908
3	B	117743	14.249218212975245
4	null	107084	12.959269622128202
5	X	80066	9.689560359767254
6	O	65645	7.944335795680082
7	A	18143	2.195659750796309
8	K	4419	0.5347858920117341
9	F	3449	0.4173968186350918
10	C	3191	0.38617374551017053

De los 20 grupos étnicos, los 4 más vulnerables son:

- H (Hispanic/Latin/Mexican): 30.75%
- W(White): 20.42%
- B (Black): 14.25%
- O (Other): 7.44%

CONCLUSIÓN

- En los últimos 4 años, desde 2020 hasta el 1 de noviembre de 2023, tuvieron lugar 826,312 delitos denunciados en la ciudad de Los Ángeles, Estados Unidos.
- La incidencia absoluta de delitos ha aumentado gradualmente desde el 2020 hasta al 2022, quedando inconclusa para el 2023, puesto que faltan los datos para los meses de noviembre y diciembre.
  - ❖ No se han identificado patrones estacionales a lo largo del año ni durante el día.
- Los delitos han sido denunciados por las víctimas o testigos de promedio a lo largo de un solo día, posterior a su incidencia.
  - ❖ Los delitos principales relacionados con abusos sexuales y malversación de fondos son los que más se han tardado en denunciar, con unas medianas superiores a un mes.
- Las áreas de la ciudad más pobladas presentan mayor incidencia de crímenes que aquellas menos pobladas.
  - ❖ El mayor número de delitos se concentra en LA Central con un total de 55,862 delitos.
  - ❖ LA Central cuenta con la mayor incidencia de delitos entre las personas sintecho de toda la ciudad; el 34.57% de los casos. No obstante, ésta representa solamente el 5.17% del total de delitos cometidos en dicha zona.
- Cerca del 54.48% de los delitos tienen lugar en las calles y las viviendas familiares.
- Los delitos principales más graves están relacionados mayoritariamente con los robos de vehículos y sus motores, así como los hurtos, y los asaltos agravados.
- En cuanto al Modus Operandi y los contextos bajo los cuales los delitos tuvieron lugar:
  - ❖ Cerca del 30% de los delitos han supuesto un robo/hurto de la propiedad de la víctima.
  - ❖ La víctima conocía a su agresor/sospechoso en el 16.26% de los casos.
  - ❖ La víctima fue agredida con el uso de cualquier arma en el 13% de los casos.
- El perfil más común de las víctimas es:
  - ❖ Aquellas en un rango de edad cerca de los 31.
  - ❖ Hombres (4.46% más que las mujeres).
  - ❖ Grupo étnico latino, el cual constituye cerca del 31% de los casos totales.[13]

[13] Se debe tener en consideración la composición demográfica de la ciudad de Los Ángeles antes de sacar conclusiones respecto al perfil de las víctimas más vulnerables. Puede ser el caso de que dicha ciudad tenga de una población significativa de origen latino relativamente a los otros grupos étnicos.