Examen ciencia de datos

Francisco Velasco Medina

Sección A

Pregunta 1: ¿Qué pruebas realizarías para garantizar la calidad de estos datos?

- La primera prueba sería ver (columna por columna) cuántos datos son nulos o NaN.

Lo que esta prueba nos indica es respecto a cuáles columnas hay que tener cuidado al utilizar datos. En caso de que omitan demasiados datos habrá que descartar la columna, imputar los datos faltantes, inferirlos a partir de otras columnas o (si son pocos) eliminar las filas con datos faltantes.

A juzgar solamente por la visualización de los datos hay tres o cuatro columnas con muchos valores faltantes.

- La segunda prueba que haría sería revisar las columnas de datos categóricos que vaya a usar (por ejemplo la columna de meses). Imprimiría los valores únicos para ver que no haya inconsistencias tipográficas.

Esta prueba permite que pueda trabajar con dichos datos posteriormente.

- La tercera prueba sería revisar que los valores estén dentro de un rango verosímil. Por ejemplo, que las coordenadas correspondan a la Ciudad de México.

Esto indicaría que los datos fueron correctamente capturados.

Pregunta 2: Identifica los delitos que van a la alza y a la baja en la CDMX (ten cuidado con los delitos con pocas ocurrencias).

Para esta pregunta agruparé los delitos según su nombre único y su año. luego imprimiré los datos para observar cuales han aumentado y disminuido. Tendré precaución con los delitos de pocas ocurrencias. Para empezar, imprimí los valores únicos de los años y los delitos (véase el código). Dado que los años presentan discontinuidades habrá que ver cuántos delitos se reportaron cada año para ver a partir de cuál año hay suficientes datos como para utilizar su información. A partir del año 2000 hay más de 100 datos por año, a partir de 2011 hay más de 1000 datos por año. Debido a la variedad de delitos que hay, creo que es prudente trabajar con los datos a partir de 2011. 2021 queda descartado porque es un año en curso.

Tras trabajar un rato con los datos llegué a dos conclusiones: trabajaré a partir de los datos de 2016, porque la cantidad de datos aumenta bastante a partir de ese año. Segundo, tendré que utilizar métricas con valores absolutos y relativos debido a las disparidades de datos reportados cada año.

Segmenté el tipo de delito por año y agregué una columna con la frecuencia relativa (ocurrencias sobre delitos totales del año)(veáse el código). Al llegar a este punto imprimí los resultados para ver las tendencias al alza y a la baja. La variedad de delitos es tan abundante que decidí concentrarme en los más comunes, así que los ordené en la tabla para analizar su evolución temporal.

	cantidad
VIOLENCIA FAMILIAR	18120
ROBO DE OBJETOS	15028
ROBO A NEGOCIO SIN VIOLENCIA	13434
DENUNCIA DE HECHOS	11373
FRAUDE	10881
AMENAZAS	9975
ROBO A TRANSEUNTE EN VIA PUBLICA CON VIOLENCIA	6319
ROBO A CASA HABITACION SIN VIOLENCIA	5755
ROBO DE VEHICULO DE SERVICIO PARTICULAR SIN VIOLENCIA	5719
FALSIFICACION DE TITULOS AL PORTADOR Y DOCUMENTOS DE CREDITO PUBLICO	4613
ROBO DE OBJETOS DEL INTERIOR DE UN VEHICULO	3857

Figura 1: Delitos más comunes de 2016.

	cantidad
VIOLENCIA FAMILIAR	28224
AMENAZAS	14386
FRAUDE	13184
ROBO DE OBJETOS	9105
ROBO A TRANSEUNTE EN VIA PUBLICA CON VIOLENCIA	8660
ROBO DE ACCESORIOS DE AUTO	7404
ROBO A NEGOCIO SIN VIOLENCIA POR FARDEROS (TIENDAS DE AUTOSERVICIO)	6160
ROBO DE VEHICULO DE SERVICIO PARTICULAR SIN VIOLENCIA	5168
ROBO DE OBJETOS DEL INTERIOR DE UN VEHICULO	4939
ROBO A NEGOCIO SIN VIOLENCIA	4450
NARCOMENUDEO POSESION SIMPLE	4213
DESPOJO	3897

Figura 2: Delitos más comunes de 2020.

Con esta información, me concentro solamente en el desarrollo de dichos delitos.

Tras observar algunas tablas (véase el código) encontré que:

Los delitos que disminuyeron son:

- Robo de objetos
- Robo a negocio sin violencia

Los delitos que aumentaron son:

- Violencia familiar
- Amenazas
- Narcomenudeo posesión simple
- Fraude (ligero aumento)
- Robo de accesorios de auto

Los delitos que subieron y bajaron son:

- Robo a transeúnte con violencia
- Robo de objetos en el interior de un vehículo

Estos últimos son delitos que iban al alza y tuvieron una gran disminución en 2020, probablemente a causa de la pandemia y la cuarentena.

Pregunta 3: ¿Cuál es la alcaldía que más delitos tiene y cuál es la que menos? ¿Por qué crees que sea esto?

Para iniciar, tengo que agrupar y contar según alcaldías y luego observar la menor y la mayor.

ABALA	1
MATIAS ROMERO	1
MARIN	1
MARIANO ESCOBEDO	1
MARAVATIO	1
...	...
XOCHICOATLAN	1
CUAUTEPEC DE HINOJOSA	1
XICO	1
SANTIAGO TULANTEPEC DE LUGO GUERRERO	2
BUENAVENTURA	2

230 rows × 1 columns

Figura 3: Alcaldías con menos crímenes.

ALVARO OBREGON	86194
BENITO JUAREZ	104735
GUSTAVO A MADERO	127220
IZTAPALAPA	190706
CUAUHTEMOC	197609

Figura 4: Alcaldías con más crímenes.

La alcaldía con más delitos es Cuauhtémoc; las alcaldías con menos delitos (y al menos un delito reportado) son 228 en total, entre las cuales están Abala, Xico, Matías Romero y Maravatio por ejemplo.

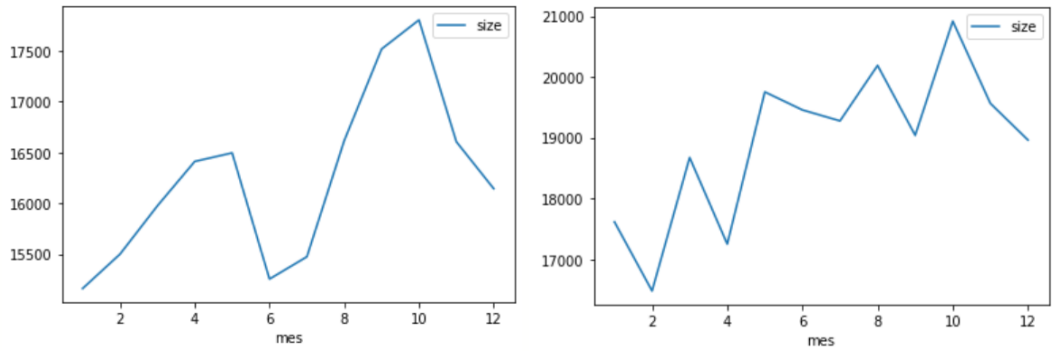
Una explicación es que Cuauhtémoc tiene mayores densidad poblacional y población total que muchas otras alcaldías y por ende ocurren más delitos de los cuales se reportan más, además de tener mayor infraestructura para el reporte de tales. Cuauhtémoc es una alcaldía por la cual mucha gente transita o trabaja.

Las alcaldías que solo reportan un delito por el contrario tienen pocos habitantes y (quizá) menor conciencia de reporte de delitos e infraestructura para reportarlos.

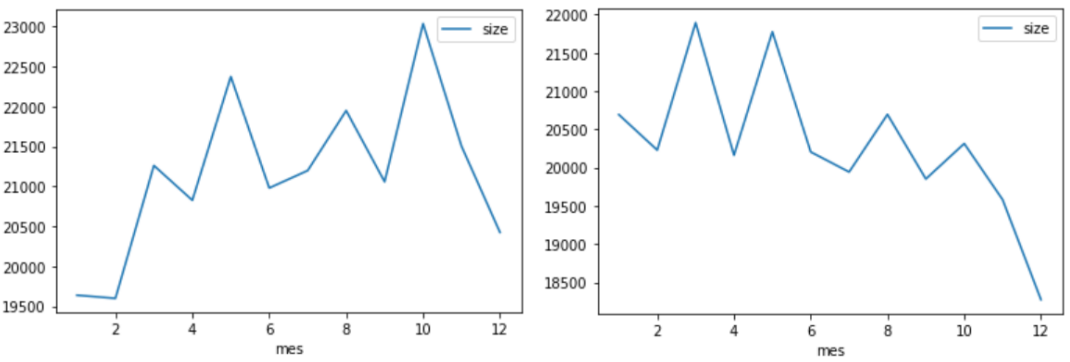
Pregunta 4: ¿Existe alguna tendencia estacional en la ocurrencia de delitos (mes, semana, día de la semana, quincenas) en la CDMX? ¿A qué crees que se deba?

Para resolver el problema es necesario segmentar según cada periodo temporal (cantidad de delitos según mes y semana, que además se deben agrupar recomendablemente según el año, y cantidad de delitos según quincena y día de la semana).

Primeramente agruparé según el año y mes (véase código). Tuve que añadir una columna con los meses representados por números para poder realizar un ordenamiento según año y luego según el mes. Me enfoqué, una vez más, en el periodo 2016-2020.



Figuras 5 y 6: Delitos por mes en 2016 y 2017.



Figuras 7 y 8: Delitos en 2018 y 2019.

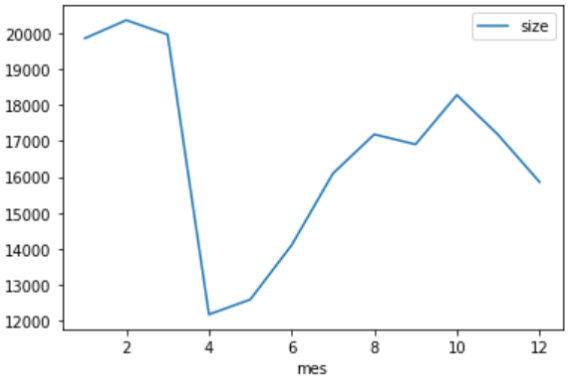


Figura 9: Delitos en 2020.

Las gráficas de mes tras mes no parecen apuntar a tendencias claras, con la excepción de que en octubre suele haber un repunte de crímenes respecto a sus meses vecinos.

La cuarentena tuvo un fortísimo impacto en la cantidad de crímenes.

Para segmentar por semana, día de la semana y quincena primero hay que convertir los datos de la columna fecha_hechos a dichos criterios respectivamente, no me alcanzó el tiempo para hacer esto. El siguiente paso a partir de aquí es graficar o hacer tablas para sacar conclusiones de las distintas medidas temporales.

Pregunta 5: ¿Cuáles son los delitos que más caracterizan a cada alcaldía? Es decir, delitos que suceden con mayor frecuencia en una alcaldía y con menor frecuencia en las demás.

Segmentaré según alcaldía y tipo de delito. Además, agregaré un contador para los delitos.

	alcaldia_hechos	delito	size
0	ABALA	DENUNCIA DE HECHOS	1
1	ACAMBARO	PRIVACION DE LA LIBERTAD PERSONAL	1
2	ACAMBARO	VIOLACION	1
3	ACAMBARO	VIOLENCIA FAMILIAR	1
4	ACAMBAY	ABUSO SEXUAL	1
5	ACAMBAY	FRAUDE	1
6	ACAMBAY	LESIONES INTENCIONALES POR ARMA DE FUEGO	1
7	ACAMBAY	VIOLENCIA FAMILIAR	1
23	ACAPULCO DE JUAREZ	PRODUCCIÓN, IMPRESIÓN, ENAJENACIÓN, DISTRIBUCI...	8
8	ACAPULCO DE JUAREZ	ABUSO SEXUAL	6
27	ACAPULCO DE JUAREZ	ROBO DE OBJETOS	6
33	ACAPULCO DE JUAREZ	VIOLACION	6
13	ACAPULCO DE JUAREZ	FRAUDE	4
18	ACAPULCO DE JUAREZ	LESIONES INTENCIONALES POR ARMA DE FUEGO	4

Figura 10: Alcaldía, delito y cantidad.

Dado que hay bastantes alcaldías y delitos, necesito restringir la selección de datos a una que se pueda analizar dentro de los límites de tiempo y sea suficientemente grande como para llegar a conclusiones interesantes. Por ende, compararé las cinco con mayor reporte de delitos y me fijaré en los quince delitos con mayor ocurrencia.

Tras observar las cinco tablas con las alcaldías y sus crímenes más reportados llegué a las siguientes conclusiones:

- Cuando las diferencias abundan encontrar similitudes puede ser revelador, cuando las similitudes abundan las diferencias pueden ser reveladoras.
- En este caso las alcaldías escogidas tienen una similitud y una diferencia clave: tienen aproximadamente los mismos crímenes con las mismas cantidades relativas, pero se distinguen en cuanto a que los crímenes se encuentran en distintos puntos de sus respectivos rankings.
- Para Cuauhtémoc e Iztapalapa el fraude es el delito con mayor ocurrencia, para Álvaro Obregón, Gustavo A Madero y Benito Juárez la violencia familiar lo es.

- Cuauhtémoc: el robo a pasajero a bordo de metro sin violencia entró dentro de los quince principales delitos, mientras que en otras alcaldías no.
- Iztapalapa: robo a repartidor apareció entre los quince principales.
- Álvaro Obregón: amenazas ocupa el segundo lugar.
- Benito Juárez: el robo de accesorios a auto se encuentra en segundo lugar, lo cual es inusualmente alto.
- Gustavo A Madero: robo a negocio tiene una posición más alta que en otras alcaldías.

Pregunta 6: Diseña un indicador que mida el nivel de "inseguridad". Genéralo al nivel de desagregación que te parezca más adecuado (ej. manzana, calle, AGEB, etc.). Analiza los resultados ¿Encontraste algún patrón interesante? ¿Qué decisiones se podrían tomar con el indicador?

Para diseñar el indicador de inseguridad hay varios factores que se pueden utilizar:

- delitos con mayores ocurrencias
- delitos según alcaldía
- alcaldías más delictivas
- áreas geoestadísticas
- estacionalidad

Las estadísticas de delitos con mayores ocurrencias en conjunto con las de delitos según alcaldía serían el primer indicador del nivel de inseguridad. Las alcaldías, manzanas y calles más delictivas serían el segundo punto del indicador y la estacionalidad el tercero (mes, semana, día de la semana y quincena).

Con el indicador se podrían tomar decisiones a nivel personal, legislativo o administrativo.

Delitos
Una persona lo podría usar para informarse de respecto a cuáles delitos debe cuidarse más y dónde ocurren con mayor frecuencia. Los legisladores, alcaldes y gobernadores lo podrían usar para saber qué se debe priorizar.

Áreas
Esta información sería útil para colocar botones de emergencia, cámaras o aumentar el patrullaje en las zonas conflictivas.

Estacionalidad
Con estos datos se pueden tomar decisiones de patrullaje y toma de precauciones alrededor de fechas conflictivas.

Para extraer información según el área utilicé la biblioteca geopandas para elaborar mapas. El código tiene errores que no alcancé a corregir. El siguiente paso a partir de aquí era analizar los mapas en búsqueda de patrones.

Sección B

Caso BOPS
¿Cuántos millones de dólares se ganaron o perdieron a causa del programa? ¿Deberían expandirse a Canadá? Explica tu razonamiento y metodología.

La pregunta uno es directa. Para ella, agrupo las ventas para ver cuánto dinero se perdió o ganó de un periodo a otro.

```
[ ]  1 g1 = df[['after', 'sales']]
     2
     3 g1 = g1.groupby(['after']).sum()
     4 g1.head()
     5
```

after	sales
0	8.046861e+07
1	6.685500e+07

Figura 11: Agrupación de ventas antes y después.

Tras convertir los números a un formato legible, el resultado es que hubo una reducción de 13,613,611 unidades en las ventas del primer al segundo

semestre desde la transición. Suponiendo que las unidades representan dólares, hubo una pérdida de trece millones seiscientos mil dólares.

¿Deberían expandirse a Canadá?

Esta cuestión es más indirecta por lo cual tengo que recolectar información para llegar a una sugerencia. Agruparé dos veces. Primero según cercanía antes y después y luego según Canadá antes y después (utilizando la información de los tres archivos).

	after	close	sales
0	0	0	4.437803e+07
1	0	1	3.609058e+07
2	1	0	3.752596e+07
3	1	1	2.932904e+07

Figura 12: Ventas (totales) antes y después, según cercanía.

usa	after	ventas
0.0	0.0	30689777.0
0.0	1.0	25853285.0
1.0	0.0	122730695.0
1.0	1.0	110455609.0

Figura 13: Ventas antes y después, según el país.

after	close	ventas
0	0	44378032.0
0	1	36090582.0
1	0	37525951.0
1	1	29329034.0

Figura 14: Ventas (brick and mortar) antes y después, según cercanía.

Las tres tablas indican una misma tendencia: sin importar el nivel de desagregación realizado, del periodo uno al periodo dos las ventas

cayeron: en EUA y Canadá, para zonas lejanas y cercanas (ya sea en total, para ventas en línea o en tiendas físicas).

Debido a que cayeron tanto en EUA (donde se implementó BOPS), como en Canadá (donde no se implementó BOPS) esto quiere decir que la caída en ventas no se le puede atribuir al inicio de BOPS.

Yo recomiendo que (a corto plazo, de uno a tres años) dejen el sistema BOPS en EUA y no lo implementen en Canadá todavía. Hay dos motivos por los cuales creo esto.

Primeramente, con esta estrategia se podría monitorear y diagnosticar más fácilmente el desempeño de BOPS, lo cual facilitaría determinar si funciona o no y los detalles pequeños específicos de cómo debe desarrollarse.

En segundo lugar, si BOPS funciona de esta manera la empresa no se perdería de algunas ganancias y si no funciona evitaría el fracaso que representaría haberlo implementado en ambos lugares. Es la estrategia más segura para ambas situaciones (las alternativas serían expandirse a Canadá inmediatamente o quitarlo de EUA inmediatamente).

La caída en ventas pudo haber sido simplemente semestral o por motivos más allá del alcance de los datos a la mano. Se necesitan datos de un mayor rango temporal y quizá más variables (ventas según tipo de productos por ejemplo y ventas de BOPS crucialmente), para diagnosticar mejor la situación. Con más datos de años pasados se podría ver si el efecto semestral es constante o si realmente fue un mal semestre consecutivo. Para juzgar a BOPS con mayor profundidad se requieren más datos a partir de 2012 (y que no lo quiten; la estrategia que propuse daría continuidad a los datos lo cual contribuiría a conseguir mayor información más pronto).

Una última recomendación para la empresa (un poco más allá de la cuestión de BOPS) se me ocurrió al leer el documento que describe el caso. En él, se relata que hay tensiones entre los departamentos de ventas en línea y ventas presenciales. Quizá deberían encontrar una buena manera de ponderar la distribución de ganancias de BOPS dado que ambos departamentos se involucran. Más aun, deberían integrarse o aprender a cooperar mejor, el documento describe que hay tensiones constantes y son reacios a cooperar.