Data Science

Procesamiento de datos con Python

Santander 2022

Proyecto Final

Aumento de las personas desaparecidas o no localizadas en México

Equipo 8

Jesús Manuel Jimenez Cardoza

Kevin Martin Rivera Castro

Luis Mateo Patricio Pineda

Minerva Estefanía Núñez Manjarrez

Nicky García Fierros

Contenido

Contenido 2

Identificación del Problema e investigación. 3

Planteamiento de preguntas. 5

Colección de Datos 6

Análisis Exploratorio de nuestro Dataset 7

Limpieza de datos y agregaciones 9

- Limpieza 9
- Agregaciones 11

Automatización y APIs 11

- Adquisición de Datos 11
- Columna dirección 12
- Automatización de proceso de peticiones 12

Transformación, filtración y ordenamiento de datos 13

- Transformación 13
- Filtración 13
- Ordenamiento 16

Siguientes pasos 17

Referencias 17

Proyecto Final - Equipo 8

Identificación del Problema e investigación.

En la actualidad, tener mayor facilidad en el acceso a las ciencias estadísticas e informáticas puede dar una visión diferente a la que se tiene. En México, la violencia y la delincuencia ocupan un espacio importante en la realidad del país (Ramírez-de-Garay & Díaz Román, 2017). La delincuencia organizada se ha convertido en un perpetrador de desapariciones (Naciones Unidas, 2022). En 58 años, México ha alcanzado la cifra de 100,000 personas desaparecidas, siendo los últimos 16 años la mayor crisis de desaparecidos (Lambertucci, 2022).

Esta realidad nos afecta a todos los mexicanos y mexicanas. Siempre hemos de comprender la situación tan difícil que es perder o extraviar a un ser querido o familiar. Las desapariciones causan impactos constantes en las familias, las cuales viven con el dolor de la pérdida y la incertidumbre. Cada día es más común ver a colectivos y grupos de búsqueda exigir ayuda al estado. Algunos colectivos incluso lideran procesos forenses que han ayudado a encontrar víctimas o dar resolución a los casos (Brewer, 2022). Es por esto que se identifica como problema de alta relevancia **el aumento en los casos de personas desaparecidas o no localizadas en México.**

Al hacer conciencia de esta impactante tragedia, en un primer acercamiento al problema nos preguntamos: ¿Cómo se investigan estos casos? ¿Cómo se visibilizan y difunden para que se resuelvan? ¿Cómo se maneja la información de tantas personas extraviadas con dignidad? ¿Existe alguna tendencia identificable en los casos? ¿Cuál es la solución a esta crisis?

A pesar de ser un problema recurrente en la actualidad mexicana, al ser un tema de carácter sensible, podemos tener una visión más objetiva al conocer los registros de las autoridades mexicanas durante un periodo de tiempo sobre las personas extraviadas o desaparecidas y que no han sido localizadas. Para poder llegar a la solución de la crisis de desaparecidos y aplicar medidas de prevención o corrección, el análisis y entendimiento de los datos forman parte importante del proceso.

Adicionalmente, el desarrollo de este proyecto también es útil para retar la información publicada por el Gobierno y ver qué tan sencillo es que los ciudadanos tengan acceso a las conclusiones a las que el Gobierno llega con los datos de acceso abierto que se generan.

Para poder dar respuesta a las preguntas pertinentes al problema, es importante definir las circunstancias de la crisis.

México, siendo un país con más de 100 mil personas desaparecidas y teniendo un sistema de búsqueda violatorio de los derechos humanos de las víctimas y de sus familias, tienen poca información del protocolo a seguir para buscar a las personas desaparecidas. Una persona se considera desaparecida cuando se desconoce su paradero sin importar el tiempo transcurrido (Flores, 2022).

Las personas no localizadas son aquellas cuya ubicación no es conocida y se considera que su ausencia no está relacionada con ningún delito. Las personas desaparecidas son aquellas cuyo paradero es desconocido por sus familiares y se presume que su ausencia se relaciona con un delito (UNAM, NA). La desaparición forzada es el arresto, secuestro, detención, o cualquier forma de privación de la libertad en

contra del consentimiento de la persona hecha por agentes del Estado o grupos aprobados por el Estado (Secretaría de Gobierno, 2016). La desaparición de personas, incluída la desaparición forzada, es una violación de los derechos humanos tanto de la persona como de los familiares de la misma. Este hecho causa un daño irreparable en la víctima y sufrimiento en los familiares (CNDH México, 2021).

El Protocolo de Actuación en caso de persona No Localizada o desaparecida dicta que si la víctima es menor de edad, la Unidad Jurídico de la entidad en donde potencialmente sucedió la desaparición debe colaborar la información y asesorar a la familia, además de dar inicio a la Alerta Amber y una carpeta de investigación. En caso de que la víctima sea mayor de edad, se tiene que avisar a la autoridad correspondiente y proporcionar los datos pertinentes para abrir la carpeta de investigación. Posteriormente la Fiscalía Especializada tomará la información recabada y continuará el proceso (UNAM, NA).

La crisis de desaparecidos desafía los recursos de las autoridades gubernamentales y exige una respuesta a una situación que día con día se convierte en un gran obstáculo que prohíbe la consolidación de una sociedad cimentada en el ejercicio de los Derechos Humanos. La Comisión Nacional de los Derechos Humanos atribuye la crisis de desaparecidos a un problema estructural derivado de la corrupción, impunidad, violencia, inseguridad y colusión del Gobierno con la delincuencia organizada que se aprovecha de la desigualdad y pobreza extrema de los mexicanos (CNDH México, 2021).

Las desapariciones forzadas son consideradas un crimen de lesa humanidad. En América Latina, este crimen se remonta a los inicios de la historia de cada país, sin embargo ha habido un incremento en la frecuencia de estos casos en los últimos 15 años. En el 2009, Raúl Lucas Lucía y Manuel Ponce Rosas, defensores de los derechos humanos, fueron desaparecidos en la región de costa chica de Guerrero y sus cuerpos fueron encontrados días después con señales de tortura. En el 2007, Edmundo Reyes Amaya y Gabriel Alberto Cruz Sánchez, militantes del Ejército Popular Revolucionario, fueron desaparecidos en la ciudad de Oaxaca. (Comisión de los Derechos Humanos del Distrito Federal, 2010)

El 26 de septiembre del 2014, 43 estudiantes de la normal rural de Ayotzinapa tomaron pacíficamente unos autobuses de transporte público del municipio de Iguala, Guerrero, para trasladarse a la Ciudad de México. Los estudiantes tenían la intención de participar en la movilización del 2 de octubre. No obstante, al caer la noche los autobuses fueron interceptados y atacados por la policía municipal de Iguala. El resultado del hecho catastrófico fueron 7 muertos, algunos de ellos con señales de tortura, y 43 estudiantes detenidos y desaparecidos. Este suceso provocó protestas e inconformidad por parte de los mexicanos hacia las autoridades. A causa de la presión ciudadana, la búsqueda de los estudiantes terminó revelando decenas de fosas clandestinas en Guerrero y estados aledaños (Gravante, 2018).

En los últimos 3 años, 58 defensores del medio ambiente y territorio mexicano fueron desaparecidos y/o asesinados. Tan solo en el 2021, 25 defensores del ambiente y territorio fueron asesinados, siendo el 41.2% actos cometidos en contra de la población indígena. Estas comunidades que denuncian actividades mineras y defienden ríos y bosques han sido señaladas como comunidades en riesgo de sufrir una agresión, desaparición y/o asesinato. En el *informe sobre la situación de las personas y comunidades defensoras de los derechos humanos ambientales en México* del 2021 se publicó un aumento del 164.44% en agresiones documentadas en contra de defensores del ambiente comparado con el 2020 (Durán Gómez, 2022).

El aumento de personas desaparecidas en México es una situación preocupante y de poca solución. La crisis ha llegado al punto en el que la alta comisionada de la ONU para los Derechos Humanos pidió a las autoridades mexicanas que incrementen los esfuerzos para acabar con las desapariciones forzadas. Se dio a conocer que de los más de 100 mil casos de desaparecidos, solamente 35 han sido condenados los perpetradores (EFE, 2022). A pesar de que existen comisiones de búsqueda en todos los estados e incluso un Centro Nacional de Identificación Humano, las soluciones a dicho problema siguen siendo escasas. Es por esto que ha surgido la creación de más de 60 colectivos y movimientos ciudadanos, como el

Movimiento por nuestros desaparecidos en México, los cuales continúan la búsqueda de la víctima cuando las autoridades no son de apoyo a la familia. Adicionalmente, tras haber logrado que la #LeyDesaparición entrará en vigor como la primera ley general en la historia de México, los colectivos trabajan para la implementación efectiva de dicha ley en cada rincón del territorio mexicano. (Movimiento por nuestros desaparecidos en México, 2021)

Una de las muchas maneras de exigir al Gobierno transparencia y eficacia en la crisis de desaparecidos es mediante la información que comparten y publican. El Registro Nacional de Datos de Personas Extraviadas o Desaparecidas (RNPED) es de acceso público y las bases de datos pueden descargarse acorde al fuero común o al fuero federal. Sin embargo, el Gobierno proporciona un dashboard para visualizar los datos. El sitio muestra que del periodo de 1964-2022, el total de personas desaparecidas, no localizadas y localizadas es de 260,561 personas, de los cuales el 40.5% son casos aún activos y el 59.5% son casos de personas localizadas (Gobierno de México, 2022).

El dashboard contiene muchas gráficas llenas de información bien digerida y estructurada, no obstante existen organizaciones que cuestionan el procesamiento de los datos y por tanto la información compartida. Data Cívica es una organización cuyo propósito es combatir la información engañosa y aportar datos estadísticos más efectivos a la realidad. Esta organización procesó los datos del RNPED y con ayuda del Centro Nacional de Búsqueda (CNB) crearon recomendaciones a implementar para el nuevo Registro Nacional de Personas Desaparecidas y No Localizadas (RNPDNO) (Data Cívica, 2019).

De igual manera, otros organismos como el CEPAD Jalisco han mostrado inconformidad con las cifras publicadas por el Gobierno de Jalisco. La acusación se dio debido a que el CEPAD corroboró los datos *raw* del RNPDNO y dio a conocer que las cifras eran diferentes a las publicadas. El Gobierno de Jalisco publicó un comunicado en el que aclaró las cifras inconsistentes y explicó que en el procesamiento de limpieza de datos hubo duplicados que no se reflejaron de inmediato en el RNPDNO, lo cual dio pie a las inconsistencias (Gobierno de Jalisco, 2022).

Planteamiento de preguntas.

Para poder llegar a una solución de un problema complejo como lo es el aumento de desaparecidos y no localizados en México, se necesita un entendimiento de los datos. Gracias a la investigación realizada, se puede inferir un nuevo cúmulo de preguntas. ¿Cómo saber cuántos desaparecidos necesitamos encontrar si la información es inconsistente o poco transparente? ¿Cómo saber que la situación está mejorando o empeorando? ¿Cómo saber si afecta más a un grupo social que a otro, y de ser así, cuál es el más vulnerable? ¿Todos los estados tienen el mismo número de desaparecidos?

Enfocando las preguntas y estructurándolas de modo que sea posible encontrar un *set* de datos que se acomode a las necesidades de la situación, las preguntas a contestar a lo largo de este proyecto son:

- ¿Existe un identificador para las personas desaparecidas?
- ¿Cuántas personas desaparecidas son encontradas?
- ¿En México solamente desaparecen mexicanos?
- ¿En qué parte del país desaparecen más personas?
- ¿Cuánto tiempo desaparecen las personas que son localizadas?
- ¿Desaparecen más hombres, mujeres o niños?
- ¿Cuál es el promedio de edad que tienen las personas desaparecidas?
- ¿Existe una relación entre el sexo de la persona desaparecida con respecto al lugar de desaparición?

- ¿En qué años han desaparecido más personas?

Con ayuda de estas preguntas es posible concluir que la base de datos que se busca debe contener información acerca del sexo, edad, lugar en el que desapareció, nacionalidad, fecha de desaparición, tiempo de desaparición, y el progreso del caso.

Colección de Datos

Un primer acercamiento al entendimiento de los datos se realiza tomando en cuenta los datos de acceso público del Registro Nacional de Datos de Personas Extraviadas o Desaparecidas (RNPED) del fuero común, que contiene datos del 2006 al 2018 (Gobierno de México, 2022). La base de datos se puede acceder aquí.

El *set* de datos fue obtenido al buscar "personas desaparecidas en México base de datos" en Google. Los primeros resultados muestran páginas que mencionan la existencia de un Registro Nacional de Datos de Personas Desaparecidas o No Localizadas. Al buscar entre las ligas, aparece el catálogo de Datos Abiertos del Gobierno de México. El catálogo tiene información acerca de la base de datos y permite la descarga del RNPED del fuero común y el fuero federal. Dado que esta base de datos fue utilizada previamente como el registro oficial de personas desaparecidas, se decidió trabajar con esta base de datos.

El gobierno de México utiliza el sistema de repositorios de datos abiertos CKAN para compartir los Datos Abiertos de las dependencias de la Administración Pública Federal, así como de gobiernos locales.

Utilizando la API de CKAN es posible consultar los detalles de los conjuntos de datos que se ofrecen en el portal datos.gob.mx/busca. Fue realizando peticiones a esta API que pudimos obtener la URL del conjunto de datos.

El endpoint para consultar datos es:

https://datos.gob.mx/busca/api/3/action/package_search?q=BUSQUEDA

Pasando el valor **"desaparecidos"** al parámetro *q* fue que obtuvimos los *datasets* relacionados. Tras asegurarnos que la petición se realizó con éxito con código HTTP 200, guardamos el formato *json* obtenido de la petición hecha a la API y observamos que tenía las llaves "help", "success" y "result", siendo esta última la llave de interés. Explorando el contenido de dicha llave concluimos que el conjunto de datos que buscamos se encuentra en "registro-nacional-de-datos-de-personas-extraviadas-o-desaparecidas-rnped". Procedimos buscando el conjunto de datos dentro del diccionario correspondiente según las investigaciones hechas y, guardamos la URL de nuestro conjunto de datos "Base de datos del RNPED del fuero común".

Una vez hecho el proceso anterior, leímos el archivo .csv a partir de la URL guardada. Anticipando que podría ocurrir un error de conexión por falla del servidor, subimos el *dataset* al repositorio de GitHub donde se encuentra nuestro cuaderno de *Jupyter* y de ésta forma, en caso de error por parte del servidor, leemos el *dataset* directamente de dicho repositorio. Tras este proceso de lectura, guardamos la versión RAW de los datos en formato csv y leemos el archivo. Obtenemos como resultado nuestro *dataframe* principal.

Análisis Exploratorio de nuestro Dataset

Primero, obtuvimos las dimensiones de nuestro *dataframe*. Encontramos que nuestro *dataframe* tiene 36265 renglones (observaciones), y 15 columnas (variables). Posteriormente, listamos cada una de las variables del *dataset* junto con su tipo de dato. Tras este proceso, pudimos observar que el tipo de dato para cada una de las variables fue *object*. Continuando con nuestra exploración, procedimos obteniendo el rango de los índices de nuestros registros. Concluimos que nuestras observaciones están indicadas por valores numéricos que comienzan en 0 y terminan en 36265.

Ya hecho este análisis estructural de nuestro *dataset*, echamos un vistazo a nuestros datos mediante las funciones "*head*" y "*tail*" para darnos una mejor idea sobre la información que tenemos a nuestra disposición para resolver nuestras preguntas. Del vistazo a los últimos 10 valores de nuestro *dataset* observamos que contiene algunos datos faltantes etiquetados por el valor NO ESPECIFICADO. Hicimos un registro de éste hecho, pues es importante tenerlo en cuenta para la limpieza de los datos.

Una vez que, mediante los pasos anteriores, nos hicimos una mejor idea de nuestros datos, exploramos las columnas. Para dicha exploración creamos una función que nos permite conocer los valores únicos de nuestro *dataset*, utilizando una estructura de datos adecuada que discrimina valores repetidos por nosotros. Las columnas exploradas y sus observaciones relevantes son:

- Fecha en que se le vio por última vez
    - En esta columna podemos observar entre qué años se encuentran los registros en nuestro *dataset* de desapariciones.
    - Encontramos que la fecha registrada más antigua es del año 1968, mientras que la más reciente es del año 2018. El hecho de que la fecha más antigua sea 1968 nos indica que hay registros de personas desaparecidas desde 1968 que, lamentablemente, hasta la fecha no han podido ser localizadas.
- País en que se le vio por última vez
    - ¿Será que existe alguna observación en el *dataset* para la cual el país donde se le vió por última vez sea distinto de México? Para contestar esta interrogante establecimos una proposición lógica que de forma automática nos dijera si la cantidad de registros que satisfacen que su valor de esta variable es mayor a cero. Dicha proposición, al ser evaluada por Python, tuvo un valor "False", por lo que concluimos que el país en el que se le vio por última vez a cada uno de nuestros registros siempre es México.
- Entidad en que se le vió por última vez
    - Los valores posibles que puede tomar la variable son los 32 estados de la República Mexicana, y "NO ESPECIFICADO".
    - De nuevo, observamos que hay NaNs en nuestros registros.
- Municipio en que se le vió por última vez
    - Notamos que fueron 1086 municipios distintos donde fueron avistadas las personas por última ocasión.
- Localidad en que se le vió por última vez
    - Notamos que fueron 2676 localidades distintas donde fueron avistadas las personas por última ocasión.
- Nacionalidad

    - Notamos que en nuestro *dataset* hay personas registradas de nacionalidad distinta a la mexicana. Para ser específicos, la variable nacionalidad toma 25 valores distintos, sin incluir “NO ESPECIFICADO”.
- Estatura
    - Nos preguntamos sobre el rango de estatura que pueden tener las personas registradas en el *dataset*.
    - Nos encontramos con un problema para contestar la pregunta anterior, en tanto que hay disparidades en las unidades en que se encuentran registradas algunas estaturas; algunos registros se encuentran en centímetros, mientras que otras se encuentran en metros. Además, identificamos otra forma en la que se registran valores nulos, esta vez como “no ESPECIFICADO”.
    - En virtud de lo anterior, contestamos esta pregunta hasta después de haber hecho una limpieza y transformación de los datos.
- Complexión
    - Pudimos observar cuatro categorías diferentes registradas en nuestro *dataset*. Además, identificamos otra forma en la que se registran valores nulos, esta vez como “No Especificado”.
- Sexo
    - Nos preguntamos si el campo permite el registro de gente no binaria.
    - Tras inspeccionar los valores únicos que toma esta variable, encontramos que dicha variable sólo admite dos valores: “HOMBRE” y “MUJER”.
- Edad
    - Nos preguntamos sobre el rango de edad de los desaparecidos.
    - Encontramos que el rango de edad va desde 1 año hasta los 103 años.
- Descripción de señas particulares
    - Observamos que esta variable contiene campos de texto de distintas longitudes.
- Etnia
    - Encontramos que hay registros que toman un valor de entre 19 etnias distintas.
    - También encontramos NaNs.
- Discapacidad
    - Encontramos que entre las personas registradas hay aquellos quienes padecen de autismo y síndrome de down.
    - También encontramos NaNs.
- Dependencia que envió la información
    - En esta columna se muestran las diferentes fiscalías y procuradurías que se encargaron de reportar los registros de desaparición.
    - En total fueron 32 distintas.

De forma muy general, del desglose anterior concluimos que nuestro *dataset* está constituido de 36265 observaciones de 15 variables indexadas del 0 al 36264. Los tipos de datos inferidos por pandas fueron en su mayoría erróneos. Con base en lo anterior, la propuesta de tipo para cada una de nuestras variables está resumida en la siguiente tabla:

Variable	Tipo
Fecha en que se le vio por ultima vez	Fecha con formato 'yyyy-mm-dd'
Hora en que se le vio por ultima vez	Hora con formato 'hh:mm:ss'
Pais en que se le vio por ultima vez	Categórico (`MEXICO`)
Entidad en que se le vio por ultima vez	Categórico (valores de `val_col_entidad`)
Municipio en que se le vio por ultima vez	Categórica (valores de `val_col_municipio`)
Localidad en que se le vio por ultima vez	Categórica (valores de `val_col_localidad`)
Nacionalidad	Categórica (valores de `val_col_nacionalidad`)
Estatura	Numérico float64 (valores en metros)
Complexion	Categórica (`DELGADA` , `MEDIANA` , `NO ESPECIFICADA` , `OBESA` , `ROBUSTA`)
Sexo	Categórica (`HOMBRE` , `MUJER`)
Edad	Numérico int64 (`1` a `103`)
Descripcion de senas particulares	Object (string)
Etnia	Categórica (valores de `val_col_etnia`)
Discapacidad	Categórica (`AUTISMO` , `NINGUNO` , `NO ESPECIFICADO` , `SINDROME DE DOWN`)
Dependencia que envio la informacion	Categórica (valores de `val_col_dependencia`)

Figura 1. Describe el tipo de información contenida por cada variable del *dataframe*.

Nuestro *dataframe* no cuenta con valores NaN estrictamente, sin embargo, tenemos registros con el valor de “NO ESPECIFICADO” y variaciones en mayúsculas y minúsculas de esta, por lo que tendremos que tomar esto en cuenta para la etapa de limpieza.

Limpieza de datos y agregaciones

Limpieza

Como nos percatamos en el análisis exploratorio, no contamos con valores faltantes indicados como NaN, sin embargo, identificamos que en nuestro *dataset* la categoría “NO ESPECIFICADO” y sus variaciones en cuanto a su escritura (“No Especificado”, no ESPECIFICADO”) fungen como registros de valores faltantes.

Realizamos entonces un conteo por columna de las ocurrencias del valor “NO ESPECIFICADO” y sus variaciones, indicando el porcentaje que representa el número de faltantes respecto al total de registros en nuestro *dataframe*. El resultado de éste proceso se resume en la siguiente tabla:

Tabla 1. Muestra la proporción de NO ESPECIFICADO contenidos en el *dataframe*.

Columna	=NO ESPECIFICADO=	Porcentaje
Fecha en que se le vio por última vez	338	0.93%
Hora en que se le vio por última vez	16	0.04%
País en que se le vio por última vez	0	0.0%
Entidad en que se le vio por última vez	29	0.08%
Municipio en que se le vio por última vez	668	1.84%
Localidad en que se le vio por última vez	3413	9.41%
Nacionalidad	2040	5.63%
Estatura	11008	30.35%
Complexión	11295	31.15%
Sexo	0	0.0%
Edad	3156	8.7%
Descripcion de senas particulares	18914	52.15%
Etnia	36133	99.64%
Discapacidad	1047	2.89%
Dependencia que envio la informacion	0	0.0%

De lo anterior, concluimos que tenemos columnas con nula presencia de faltantes, poca presencia y algunas columnas en donde la cantidad de faltantes es considerable.

Nuestro proceso de limpieza fue como a continuación:

- Creamos una copia de nuestro *dataframe* original.
- Transformamos todos los valores de “NO ESPECIFICADO” y sus variaciones a NaN
    - Tras este proceso de transformación, verificamos que la cantidad de NaNs coincidiera con los de “NO ESPECIFICADO” y sus variaciones ya contadas antes.
- Limpiamos por columna los NaNs, discriminando según sea adecuado.
    - Describimos el proceso de cada columna por separado en el cuaderno de jupyter, pues requiere de un análisis cuidadoso según sea el caso particular de cada campo.
- Registramos aquellas columnas donde mantener los valores de “NO ESPECIFICADO” sea adecuado y transformamos los NaN a “NO ESPECIFICADO” para estas columnas particulares.

    - Las columnas en cuestión fueron las siguientes:
        - Municipio en que se le vio por última vez
        - Localidad en que se le vio por última vez
        - Nacionalidad
        - Complexión
        - Descripcion de senas particulares
        - Discapacidad
- Verificamos que nuestro *dataframe* esté libre de datos faltantes.
- Comparamos las dimensiones del *dataframe* original con el limpio.
    - Se eliminaron 364 registros, lo que representa un 1.004% del total original.
- Modificamos los índices de los registros
    - Modificamos el valor inicial para que los registros empiecen desde 1 en lugar de 0. De esta forma, el número de filas ya no coincide con el índice del último registro.
- Modificamos el nombre de las columnas para que estén en la convención estándar *snake case*.

Agregaciones

Utilizamos la agregación *describe* sobre nuestro *dataset* para obtener de forma organizada por columna con valores numéricos (estatura y edad):

- La cantidad de observaciones no nulas
- Los valores máximos
- Los valores mínimos
- La media
- La desviación estándar
- Los cuartiles 25%, 50%, y 75%

Encontramos que la estatura media es de 1.639 m con desviación estándar 0.1360. El 50% de los registros tienen una altura menor a 1.64 m. Más aún, contestamos la incógnita sobre el rango de alturas: resultó que el rango de alturas es de 0.3 m a 2.04 m.
Por otro lado, la edad media es de 30.5499 años, con desviación estándar de 14.37 años, y el 50% de las personas registradas como desaparecidas, en nuestro *dataset*, tienen una edad menor o igual a 28 años.

Automatización y APIs

Adquisición de Datos

Como ya se mencionó anteriormente en la sección de *Colección de Datos* del presente documento, utilizamos la API de CKAN para consultar los detalles de los conjuntos de datos que se ofrecen en el portal datos.gob.mx/busca.

Además, utilizamos una API de geolocalización que, a partir de una dirección, nos regresa las coordenadas en valores de latitud y longitud. La API que usaremos es proporcionada por Positionstack, cuyo plan gratuito nos permite realizar hasta 25 mil consultas por mes. La API requiere de una API access key. En virtud de la restricción sobre la cantidad de consultas, obtuvimos dos tokens para cubrir nuestro conjunto de datos.

Columna dirección

Creamos una nueva columna temporal con el nombre de "dirección", cuyo propósito es almacenar el valor del nombre del municipio junto con el nombre del país separados por coma. Fue necesario generar la columna con este formato porque nos permite obtener mayor precisión en las coordenadas.

Recordando que en la columna de municipio mantuvimos la categoría NO ESPECIFICADO, tenemos que buscar una alternativa para formar nuestro valor de dirección, en estos casos, se tomó la decisión de usar el valor de la columna entidad para completar el valor de la dirección y para que la API nos pueda regresar un resultado.

Automatización de proceso de peticiones

Con base en la documentación oficial de la API, definimos los valores de nuestro endpoint y nuestro diccionario de parámetros. Dividimos nuestro conjunto en dos partes, pues por los límites en número de consultas requerimos hacerlos con dos tokens distintos.

Automatizamos el proceso de peticiones mediante ciclos for, iterando sobre cada elemento de la columna direcciones. En resumen, el proceso de automatización se divide en los siguientes subprocesos:

- Definimos la query que toma la petición a la API.
- Utilizamos el método GET de *requests* para realizar la petición a nuestro endpoint y enviar los parámetros como información extra que la API necesita para validar dicha petición.
- Guardamos el JSON, y obtenemos de este los valores de longitud y latitud. Dicha información la guardamos en una cadena, donde los valores están separados por una coma.
- Eliminamos la columna temporal de dirección.

Para manejar los posibles errores que pudieran ocurrir de los procesos anteriores utilizamos bloques *"try-except"*. En caso de que ocurra algún error, almacenamos una tupla "(0,0)" en lugar de las coordenadas. En total obtuvimos 35901 datos.

Finalmente, creamos una columna llamada "coordenadas" en nuestro *dataframe* con los valores obtenidos del proceso en párrafos anteriores. Para esto, dentro de un bloque *"try-except"* convertimos la lista donde almacenamos las coordenadas en un dataframe para después convertirlo en un *csv* para consultas posteriores en caso de que no se pudieran hacer las peticiones por cuestiones de tiempo o por errores de parte del servicio.

Transformación, filtración y ordenamiento de datos

Transformación

Nuestro proceso de transformación se resume en los siguientes subprocesos:

- Casting
    - Observamos nuestras columnas y su tipo de dato.
    - A pesar de haber realizado el casting correspondiente en nuestras variables numéricas, observamos que aún tenemos dos columnas que poseen un tipo de dato incorrecto: fecha y hora.
    - Para asignar el tipo de dato correcto a las columnas con tipo de dato incorrecto utilizamos un diccionario de conversión, corrigiendo detalles como que al hacer el casting correspondiente a los valores de la columna se agregó información de más por la forma en que funcionan los métodos de pandas empleados.

- Manipulación de Strings
    - Observamos que las cadenas de nuestras variables de tipo string están en mayúsculas.
    - Modificamos los registros en las columnas con tipo string de modo que sólo la primera letra de cada palabra comience con mayúscula, manteniendo las mayúsculas de las siglas de las dependencias.

Filtración

Tras el proceso de transformación, creamos nuevas columnas y segmentamos el *dataframe* para poder contestar las preguntas planteadas al inicio. En resumen, el proceso de filtración se llevó a cabo en los siguientes puntos principales, guiados por nuestras preguntas:

- **¿En qué parte del país desaparecen más personas?**
    - Se realizó una agrupación por entidad para contestar esta pregunta.
    - Concluímos que de 1968 a 2018 hubo más desapariciones en el estado de Tamaulipas, con un total de 5,987.
    - Adicional a esta pregunta surgió otra muy natural: **¿cuál es el estado con menor cantidad de desapariciones?**
        - Del mismo *dataframe* generado por la agrupación pudimos contestar esta pregunta, cuya respuesta es Tlaxcala con 24 desapariciones.
    - Aprovechando la naturaleza del registro, pudimos complementar la respuesta anterior, incluyendo el municipio y la localidad.
    - Encontramos que la localidad de Matamoros, Matamoros Tamaulipas tiene más registros de desapariciones con un total de 1157. Más aún, los primeros cinco lugares son ocupados por Tamaulipas principalmente, Nuevo León y Puebla, mientras que los últimos cinco lugares por Tamaulipas y Puebla.

- ¿En México solamente desaparecen mexicanos?
    - Para contestar esta pregunta hicimos uso de una agrupación por el campo de “nacionalidad”.

    - Apreciamos que existen 2,027 casos no especificados.
    - En su mayoría las personas registradas tienen nacionalidad mexicana, no obstante, hay mención de 24 nacionalidades distintas a la mexicana, siendo la estadounidense la segunda con mayor ocurrencia.

- ¿Desaparecen más hombres, mujeres o niños?
    - En virtud de que no existe una columna que especifica si la persona registrada es o no mayor de edad, contestamos la pregunta por partes.
    - Primero, mediante una agrupación, observamos los registros por sexo. Según los datos encontrados, hay 9,226 mujeres desaparecidas y 26,675 hombres desaparecidos. Por consiguiente, la mayoría de las personas desaparecidas en México están registradas como hombres.
    - Con ayuda de una agrupación por sexo y edad, observamos que los hombres de 28, 26 y 21 años de edad, en ese órden, son los que aparecen con mayor frecuencia en el *dataset*; mientras que las mujeres de 28 y 16 años son las más vulnerables en la crisis de personas desaparecidas. Las personas de la tercera edad son las que menos registros de desaparición tienen.
    - Para conocer la proporción de menores de edad que desaparecen a comparación de aquellos que son mayores de edad generamos un diccionario en donde se evalúa si la edad registrada es mayor-igual o menor a 18. Con este diccionario hicimos una consulta al *dataframe* y encontramos que hay 29466 casos de mayores de edad desaparecidos, siendo esto el 82% del total (36,901) de los registros de personas desaparecidas.
    - En conclusión, desaparecen más hombres que mujeres en México. Adicionalmente, desaparecen más personas mayores de edad que menores de edad.

- ¿Existe una relación entre el sexo de la persona desaparecida con respecto al lugar de desaparición?
    - Con la información que ofrece el registro no se puede afirmar que exista una relación directa entre el lugar de desaparición y el sexo de la persona. Sin embargo, se puede visualizar cuál es la frecuencia de edad más común de las personas según su sexo y lugar de desaparición.
    - Para lo anterior, hicimos una agrupación por sexo, mostrando la edad y la entidad.
    - Del resultado de la agrupación, la edad tanto de hombres como mujeres con más casos de desaparecidos es de 28 años. No obstante, para los hombres el lugar de la desaparición es en Tamaulipas. Las mujeres tienen más casos de desaparición en el Estado de México.
    - Lo mencionado anteriormente va acorde a la investigación realizada en el proyecto, dado que existe un aumento de violencia a las mujeres en el Estado de México. De igual manera, en el norte del país existe mucha violencia de grupos armados liderados en su mayoría por hombres.
    - Concluimos que, a pesar de que **esto no es una confirmación absoluta** de que existe una relación entre la edad, **el sexo, y el lugar de desaparición, podría llegar a ser un dato que respalde una investigación formal al respecto.**

- ¿Cuántas personas con discapacidad han desaparecido en México?
    - Mientras nos encontrábamos realizando la investigación fue que surgió esta pregunta. Nos pareció que es una pregunta relevante una vez que exploramos el contenido del

registro, pues es común que en México no se tome en consideración a las personas con capacidades diferentes; sin embargo, recordemos que la realidad del país también les afecta a ellas.
    - Mediante una agrupación en la variable “discapacidad” encontramos que la mayoría de las personas desaparecidas no tienen una discapacidad. Hay 1043 personas que no tienen una discapacidad especificada, y existen 15 personas desaparecidas en el registro, de las cuales 7 de ellas tienen autismo, y 8 presentan Síndrome de Down.

- ¿Cuánto tiempo desaparecen las personas localizadas?
    - El registro usado en este proyecto no cuenta con el estado del caso. Es por esto que esta pregunta no se puede responder en su totalidad. Sin embargo, saber cuántos días lleva la persona desaparecida es información relevante.

- ¿Cuánto tiempo llevan desaparecidas las personas registradas en el conjunto?
    - En virtud de que el registro no contiene datos al día de hoy, sería incorrecto asumir que las personas siguen desaparecidas en el presente. En la información del catálogo de datos abiertos del Gobierno, las cifras presentan el total de registros de personas relacionadas con averiguaciones previas, carpetas de investigación o actas circunstanciadas del fuero común que permanecen sin localizar al corte del 30 de abril del 2018 distribuidas por año. Por tanto, tomaremos como última fecha el 30 de abril del 2018.
    - Al obtener el máximo valor de la columna fecha, obtuvimos que el último registro es del 29 de abril del año 2018.
    - Con una función *helper* que cuenta los días que han transcurrido dada una fecha hasta el 30 de abril del 2018, aplicamos un *apply* que nos ayuda a generar un *dataframe* que nos muestre la cantidad de días que han transcurrido desde el día de su registro hasta el 30 de abril del 2018.

- ¿En qué años han desaparecido más personas?
    - Para contestar esta pregunta únicamente con el año, requerimos extraer el año de la fecha y agregar la variable a nuestro *dataframe*.
    - Con la columna agregada, hicimos una agrupación por año y ordenamos por total de forma descendente.
    - Encontramos que el año con el mayor personas desaparecidas fue el 2017 con 5,426 registros. Esta información respalda la investigación preliminar del problema. Podemos notar que los primeros 13 lugares de la lista son años del 2006 en adelante, justo cuando inicia la crisis de desaparecidos y el aumento más rápido de casos de personas desaparecidas en la historia de México (desde que se inició el registro de las personas desaparecidas).

Es importante mencionar que hubo dos incógnitas planteadas que no pudimos resolver con el *dataset*. El motivo de que no sea posible dar solución a las preguntas “¿cuántas personas desaparecidas son encontradas?” y “¿existe un identificador para las personas desaparecidas?” radica en que en el registro no existe una columna que tenga el estado del caso o la carpeta de investigación. Más aún, no existe un identificador para las personas desaparecidas, de modo que no tenemos forma de buscar la información relevante para contestar estas preguntas en otros lugares. Cabe mencionar que, de existir un identificador

para las personas desaparecidas, podríamos utilizar dicho identificador como llave primaria en análisis más complejos de los datos presentados.

Ordenamiento

Previo a guardar nuestro *dataframe* realizamos un proceso de reordenamiento con el propósito de que nuestros datos sean más legibles. El resultado del ordenamiento de los campos se ve reflejada en la siguiente lista ordenada:

1. anio
2. fecha
3. hora
4. dias_desaparecido
5. entidad
6. municipio
7. localidad
8. coordenadas
9. nacionalidad
10. sexo
11. edad
12. estatura
13. complexion
14. senas_particulares
15. discapacidad
16. dependencia_origen

Por último, exportamos el *dataframe* resultante a formato *csv*. Este *dataframe* puede ser descargado en la siguiente liga:

Proyecto-de-Python-BEDU/RNPEDFC_Final.csv at main · MinervaNunez/Proyecto-de-Python-BEDU (github.com)

Cabe destacar que el análisis completo con lujo de detalle se encuentra disponible en el repositorio del proyecto, al cual se puede acceder desde la siguiente dirección electrónica: MinervaNunez/Proyecto-de-Python-BEDU: Proyecto de Python para el curso de "Ciencia de Datos" impartido por BEDU a través de Becas Santander. (github.com)

Siguientes pasos

La crisis de personas desaparecidas en México es un problema que tiene varios niveles de complejidad e involucra muchos factores como la corrupción, la pobreza, la educación, etc. Dicho problema no puede ser resuelto con solo ver los datos, requiere de una solución transversal. No obstante, la exploración y entendimiento de los datos proveen nuevas preguntas e iniciativas para llegar a la solución de manera más rápida.

En México desaparecen más personas en el norte del país, siendo Tamaulipas el estado con más casos de personas desaparecidas. Los hombres en sus veintes son las víctimas mayoritarias de esta crisis, no obstante también hay casos de mujeres, menores de edad, y personas con discapacidad. Los datos muestran el aumento en los casos de personas desaparecidas año con año, siendo el 2017 el año con más casos de personas desaparecidas registradas en el *dataset*.

El análisis de los datos no solo resuelve las preguntas planteadas, sino también da ideas para la mejora de la colección de datos. Parte importante de la información registrada en una base de datos es tener una columna de valores únicos por registro, para así poder tener identificadores definidos por caso. Además, esto puede fungir como llave primaria para el manejo de los datos en análisis más extensos. De igual forma, se debe recabar más información acerca de la etnia de las personas desaparecidas, ya que es un dato importante para saber si existe un ataque a las comunidades indígenas. Por último, hay datos que deben ser recabados de forma obligatoria, como lo es la entidad en la que se le vio por última vez. No solo es información clave para su búsqueda, sino también es un estadístico importante para la planificación de estrategias de resolución.

Al haber limpiado la información, llegamos a tener una visión un poco más comprensible para las preguntas realizadas. Los siguientes pasos propuestos son un dashboard con gráficas para tener una visión de la información más gráfica y así tener una población más consciente e informada de la crisis que va en aumento en México. Con base al set de datos se podría exportar *dataframes* de particulares observaciones para crear estudios estadísticos con respecto a los años, entidades y sexo de las personas desaparecidas y se podrían crear hipótesis políticas; por ejemplo, con las coordenadas obtenidas mediante peticiones a APIs, se podría analizar la posibilidad de proponer estrategias para mitigar futuras desapariciones.

En conclusión, el aumento de personas desaparecidas en México es una situación que nos involucra y nos concierne a todas y todos los mexicanos. Parte importante de ser un ciudadano involucrado de forma activa en el país es el informarse, cuestionar, y verificar la información que proviene de fuentes oficiales. El análisis, la limpieza, y la manipulación de datos ayudan a entender mejor el problema y a proponer estrategias y modificaciones a las soluciones actuales. Cada paso que se toma para deshacer la complejidad del problema es crucial para salvar más vidas sin importar la extensión del análisis. El conocimiento no solo es poder, sino también es vital.

Referencias

Naciones Unidas. (12 de 04 de 2022). *México: La prevención debe ser central en la política nacional para detener las desapariciones forzadas, señala Comité de la ONU*. Obtenido de Naciones Unidas: https://www.ohchr.org/es/press-releases/2022/04/mexico-prevention-must-be-central-national-policy-stop-enforced

Lambertucci, C. (16 de 05 de 2022). *México supera las 100.000 personas desaparecidas*. Obtenido de El País: https://elpais.com/mexico/2022-05-17/mexico-supera-las-100000-personas-desparecidas.html#?prm=copy_link

UNAM. (NA). *PROTOCOLO DE ACTUACIÓN EN CASO DE PERSONA NO LOCALIZADA, PARTE DE LA COMUNIDAD UNIVERSITARIA*. Obtenido de Servicio a la Comunidad UNAM: http://www.serviciosalacomunidad.unam.mx/index_htm_files/Protocolo_de_persona_no_localizada_parte_de_la_comunidad_universitaria.pdf

Secretaría de Gobierno. (22 de 12 de 2016). *¿Qué es la desaparición forzada?* Obtenido de Gobierno de México: https://www.gob.mx/segob/articulos/que-es-la-desaparicion-forzada?idiom=es

CNDH México. (2021). *PERSONAS DESAPARECIDAS*. Obtenido de Informe de Actividades 2021: http://informe.cndh.org.mx/menu.aspx?id=30062

Flores, S. (22 de 05 de 2022). *Qué hacer para buscar a un desaparecido: una guía a la que nadie debería tener que recurrir*. Obtenido de Animal Político: https://www.animalpolitico.com/2022/05/como-buscar-a-un-desaparecido-guia-nadie-deberia-necesitar/

Brewer, S. (16 de 05 de 2022). *México: 100.000 personas desaparecidas y no localizadas*. Obtenido de WOLA: https://www.wola.org/es/analisis/mexico-personas-desaparecidas-y-no-localizadas/

CNDH México. (2021). *PERSONAS DESAPARECIDAS Y NO LOCALIZADAS*. Obtenido de Informe de Actividades 2021: http://informe.cndh.org.mx/menu.aspx?id=50062

Comisión de los Derechos Humanos del Distrito Federal. (2010). *Defensor.* Ciudad de México: CDHDF.

Gravante, T. (2018). Desaparición forzada y trauma cultural en México: el movimiento de Ayotzinapa. *Convergencia, 25*(77), 13-28.

Durán Gómez, T. (31 de 03 de 2022). *México: 58 defensores de ambiente y territorio fueron asesinados en los últimos tres años* . Obtenido de MONGABAY: Periodismo Ambiental Independiente en Latinoamérica: https://es.mongabay.com/2022/03/mexico-58-defensores-de-ambiente-y-territorio-asesinados/

EFE. (17 de 05 de 2022). *ONU pide a México aumentar esfuerzos contra desapariciones forzadas* . Obtenido de El Financiero:

https://www.elfinanciero.com.mx/mundo/2022/05/17/onu-pide-a-mexico-aumentar-esfuerzos-contra-desapariciones-forzadas/

Movimiento por nuestros desaparecidos en México. (2021). *LA LEY DESAPARICIÓN Y LA IMPLEMENTACIÓN*. Obtenido de Movimiento por nuestros desaparecidos en México: https://memoriamndm.org/ley-desaparicion/

Gobierno de México. (2022). *Contexto General*. Obtenido de Gobierno de México : https://versionpublicarnpdno.segob.gob.mx/Dashboard/ContextoGeneral

Data Cívica. (2019). *Evaluación para el diseño del Nuevo Registro Nacional de Personas Desaparecidas*. Obtenido de Data Cívica: https://registros-desaparecidos.datacivica.org

Gobierno de Jalisco. (29 de 03 de 2022). *ACLARACIÓN SOBRE LAS CIFRAS DE JALISCO EN EL REGISTRO NACIONAL DE PERSONAS DESAPARECIDAS Y NO LOCALIZADAS*. Obtenido de Jalisco: Gobierno del Estado: https://www.jalisco.gob.mx/en/prensa/noticias/141351

Ramírez-de-Garay, D. &. (2017). Los efectos de la política de prevención del crímen y la violencia en México. *Revista CIDOB d'Afers Internacionals*, 101-128.

Gobierno de México. (2022). *Base de datos del RNPED del fuero común Descargar .* Obtenido de Datos Abiertos: https://datos.gob.mx/busca/dataset/registro-nacional-de-datos-de-personas-extraviadas-o-desaparecidas-rnped/resource/3042fa5b-5635-4575-ab75-a1057a2b2481