# UNIVERSIDAD INTERNACIONAL DE LA RIOJA

## MAESTRÍA EN ANÁLISIS Y VISUALIZACIÓN DE DATOS

Métodos de captura y almacenamiento de datos

## Actividad 1: Limpieza de dataset

Presenta

**Alain Iglesias Herrera**

**829826**

Bárbaro Jorge Ferro Castro

CDMX, mayo 2022

| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

Tabla de contenidos

| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

## Introducción

El presente trabajo tiene como objetivo identificar adecuadamente los errores más comunes que se pueden encontrar en un conjunto de datos, prepararlos y realizar mejoras para optimizar y obtener mejores resultados a la hora de analizarlos. El enfoque del mismo es hacia los **tipos de problemas** que serán identificados a través de la búsqueda de patrones entre los problemas introducidos dentro del dataset proporcionado. Para la identificación de estos tipos de errores nos apoyaremos en la herramienta **Jupyter** de Microsoft y la librería **Pandas** de Python.

| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

# Desarrollo

## Identificación de problemas

1. El primer tipo de problema que se detectó en el dataset está relacionado con **valores nulos**. Estos valores no aportan nada valioso al conjunto de datos y pueden ser eliminados y/o modificados por un valor por default u otro que nos acomode. Lo cual se puede ver reflejado en la columna **Range** que tiene todos sus campos vacíos, la columna **City** tiene **321** campos nulos y **State** tiene 3. En este caso es posible eliminar la columna **Range** completamente. Si los valores nulos de **City** tienen un **State** válido es posible sustituir ese valor nulo por la ciudad de **San Francisco** que es la ciudad que se encuentra en la muestra, lo mismo sucedería con el **State,** si tiene un City válido se puede inferir.

```
CrimeId                0
OriginalCrimeTypeName  0
OffenseDate            0
CallTime               0
CallDateTime           0
Disposition            0
Address                0
City                 321
State                  3
AgencyId               0
Range              10051
AddressType            0
dtype: int64
```

2. El segundo tipo de problema está relacionado con **errores de ortografía.** Se puede apreciar en diferentes campos cómo existen valores mal escritos que causan confusión. Por ejemplo, en el campo **AddressType** existe un campo con valor **Intersectioon.** La corrección de estos errores se puede corregir realizando un reemplazo a toda la columna con el nombre correctamente escrito.

| | CrimeId | OriginalCrimeTypeName | OffenseDate | CallTime | CallDateTime | Disposition | Address | City | State | AgencyId | Range | AddressType |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3775 | 160932590 | Traffic Stop | 2016-04-02T00:00:00 | 17:07 | 2016-04-02T17:07:00 | ADV | 24th St/capp St | San Francisco | CA | 1 | NaN | Intersectioon |

3. El tercer tipo de problema va de la mano con el anterior y tiene que ver con **valores inconsistentes**. Esto se puede encontrar en diferentes columnas del conjunto de datos. Por ejemplo, el campo de **AgencyId** contiene 3 filas con el valor de **CA** en lugar del identificador de la agencia y **City** tiene un valor de **S** en lugar de la ciudad. Para solucionarlo en este caso particular, el campo **AgencyId** se pondría igual a 1 dado que este es el id que predomina en el dataset y para el campo **City** a la ciudad que predomina igualmente.

| | CrimeId | OriginalCrimeTypeName | OffenseDate | CallTime | CallDateTime | Disposition | Address | City | State | AgencyId | Range | AddressType |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5771 | 160942112 | Auto Boost / Strip | 2016-04-03T00:00:00 | 14:30 | 2016-04-03T14:30:00 | REP | Martin Luther King Dr/bowling Green Dr | NaN | NaN | CA | NaN | 1 |
| 8021 | 160952280 | Auto Boost / Strip | 2016-04-04T00:00:00 | 14:46 | 2016-04-04T14:46:00 | REP | Martin Luther King Dr/nancy Pelosi Dr | S | NaN | CA | NaN | 1 |
| 8473 | 160953118 | Auto Boost / Strip | 2016-04-04T00:00:00 | 18:11 | 2016-04-04T18:11:00 | REP | Conservatory Drive E/john F Kennedy Dr | NaN | NaN | CA | NaN | 1 |

En el caso de la columna **OriginalCrimeTypeName**, a nuestro modo de ver, la más importante, contiene una gran cantidad de valores numéricos que no permiten conocer el nombre del crimen que se está cometiendo, lo cual le quita el sentido al

objetivo de este dataset, ya que el nombre de la columna hace referencia a un nombre de crimen no a un código que es lo que se pudiera interpretar de esos valores. Para darle una solución a este problema se requiere más contexto de modo que se pueda discernir si se trata de una codificación o de error humano.

4. El cuarto tipo de problema esta relacionado con **identificadores duplicados**. Existen 2 identificadores (ID's) duplicados en toda la serie, los ID's son: **160913455** y **160950496**, ambos se repiten 3 veces para un total de 6. Los ID's permiten que podamos identificar cada fila como única entre todo el conjunto de datos, por lo que tener varias filas con el mismo ID causaría inconsistencia a la hora de procesar el dataset. En este caso la solución más factible es aumentar en 1 el último dígito del ID (verificando previamente que no colisione con otros ID's), dado que generalmente es un número consecutivo el que se emplea para este tipo de dato.

```
26      160913455
1707    160913455
3792    160913455
7045    160950496
7046    160950496
7047    160950496
Name: CrimeId, dtype: int64
```

5. El quinto tipo de problema esta relacionado con **columnas que no aportan información de calidad** al conjunto de datos. La columna **AgencyId** en este caso contiene información que no es relevante para el dataset, su valor es el ID que debería tener en otra tabla para identificar de cuál agencia estamos hablando. Eliminar esta columna sería la solución más adecuada ya que no aporta ningún dato relevante para futuros análisis.

| AgencyId |
|---|
| 1 |
| 1 |
| 1 |

6. El sexto tipo de problema identificado es relativo al **manejo de fechas y horas** en el dataset. El campo **OffenseDate** contiene un valor de tipo datetime, o sea, fecha y hora, pero solo la fecha es relevante dado que la hora no es registrada. Existe un dato en esta columna con fecha futura "2025-04-04T00:00:00", que debe ser actualizado al año que se está manejando: 2016. Para este campo se realizaría una actualización para dejar en su valor solo la fecha que es el dato relevante. Por otro lado, tenemos los campos **CallTime** y **CallDateTime** cuya situación es bien similar, solo que en este caso tenemos que se está registrando la hora en ambos campos. La solución más adecuada sería eliminar la columna **CallTime** ya que de esta forma se

| OffenseDate |
|---|
| 2016-03-30T00:00:00 |
| 2016-03-31T00:00:00 |
| 2016-03-31T00:00:00 |

podrían tener todos los datos en un mismo campo y se eliminarían los datos redundantes.

| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

**Metodología**

Se recomienda realizar el proceso de limpieza del dataset en un **Jupyter notebook**, esta herramienta ayudará mucho a ir visualizando los resultados obtenidos tras aplicar las diferentes acciones para lograr nuestro objetivo.

1. Primeramente, se debe hacer un bosquejo general de una muestra del conjunto de datos (unos 20 rows estarán bien para familiarizarse con la estructura y datos que contenga la base)
2. Posteriormente, se prestará atención a los errores más fáciles de detectar y corregir, como faltas ortográficas, valores nulos, inconsistentes, duplicados, etc.
3. Revisar que todas las columnas aporten información relevante, en caso contrario se eliminarán, ex. **Range y AgencyId**
4. Verificar que no existan datos duplicados entre diferentes columnas, en ocasiones se pueden incluir datos que están presentes en una columna como un subconjunto de otra, ex. **CallTime** y **CallDateTime**.

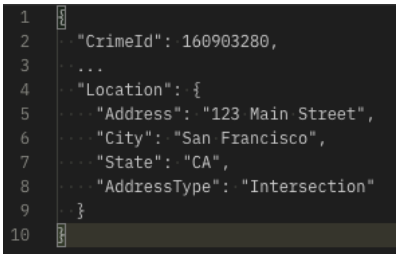
Al seguir estos pasos se podrá afirmar que este catálogo de datos está limpio y listo para su análisis.

| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

**Propuesta de mejoras en el conjunto de datos**

Las mejoras introducidas pueden ser variadas dependiendo del objetivo para el que se requieran los datos. Para la segunda propuesta se asume que se aplicaría sobre el conjunto de datos en **formato json**.

- La primera propuesta ya fue identificada en el tipo de problema número 5, y es referida a eliminar la columna **CallTime** ya que esta solo contiene datos redundantes.
- La segunda propuesta estaría enfocada a agrupar los campos referidos a la dirección, por ejemplo, un sub-objeto con la estructura mostrada en la imagen. Para lograr este agrupamiento se creó una función que realiza la modificación automáticamente. Se adjunta link de repositorio en Github: https://github.com/aiherrera/MCAD-tasks



| Asignatura | Datos del alumno | Fecha |
|---|---|---|
| Métodos de captura y almacenamiento de datos | Apellidos: Iglesias Herrera | 06 de junio de 2022 |
| | Nombre: Alain | |

## Conclusiones

Detectar áreas de mejora en los datos con los que se van a trabajar es una ventaja para mantener una base de datos con un alto grado de calidad y consistencia. La realización de este ejercicio permitió sentar bases sobre las que se creó una metodología que sirvió como guía para la limpieza de este dataset y que será útil para la preparación de otros con errores semejantes.

La propuesta de mejoras sobre los datos que se trabajaron permitió ver más allá de la identificación de los errores, pues sirvió para ejercitar las habilidades de investigación, así como, el desarrollo de la capacidad de encontrar estrategias para el procesamiento de la información.