 **Edwin Giovanny Cuenca Moreno**

**Gestión de Datos**

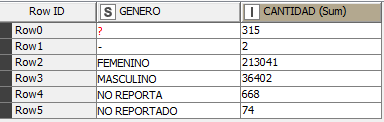
**Maestría en IoT**

**TALLER 1**

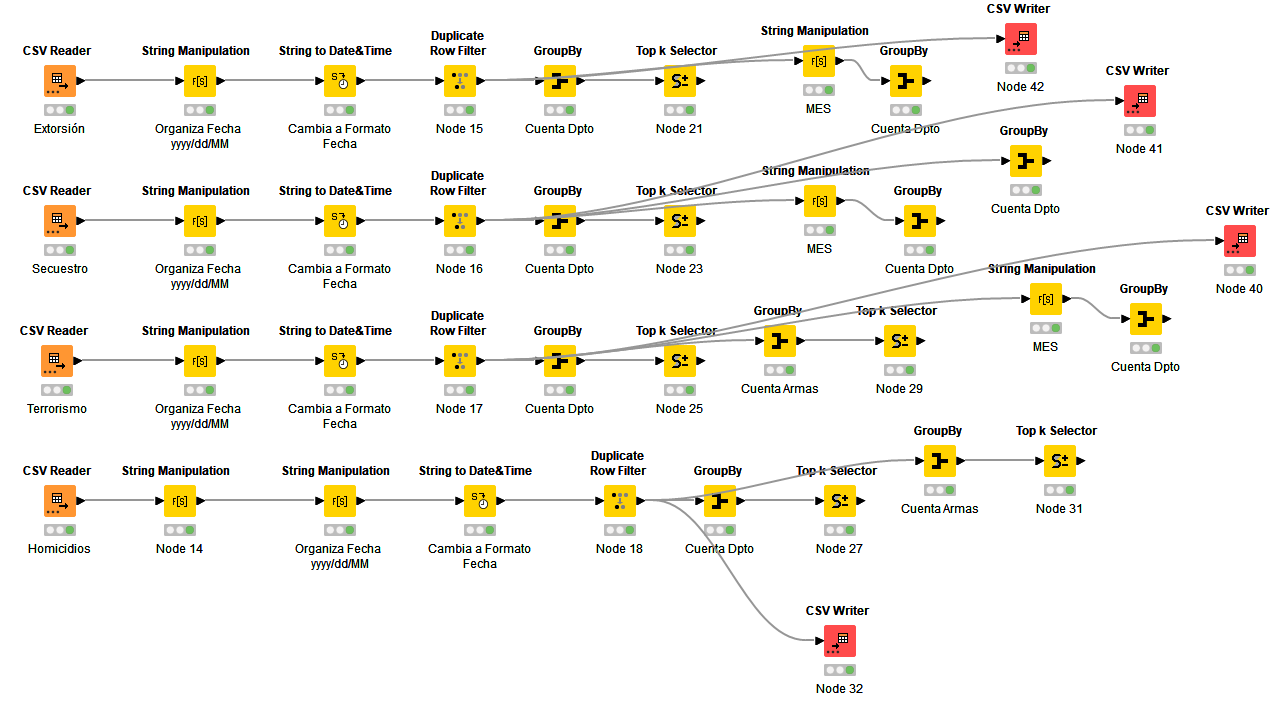
1. Seleccione al menos 4 conjuntos de datos y realice un perfilamiento de cada uno. Describa los hallazgos encontrados en cada conjunto de datos. No olvide mencionar su estructura y aspectos relevantes de calidad como campos nulos, departamentos mal escritos, formatos de fechas incorrectos, entre otros. Si no evidencia ningún problema de calidad de datos, también menciónelo.

Dentro del conjunto de datos suministrado, se realizó la selección de los siguientes 4 dataset para realizar el perfilamiento de cada uno de estos:

* **Extorsión**
* Este archivo contiene información de la cantidad de eventos de extorsión registrados entre el año 1996 y 2022, discriminados por departamento y municipio cada uno de estos con su respectivo código de identificación.
* Dentro de las características de las variables se evidencia que la variable de fecha es de tipo string en formato dd/MM/yyyy hh:hh, para lo cual se realizó el respectivo ajuste al formato de fecha al “MM/dd/yyyy” mediante un bloque de String Manipulation para extraer solo la fecha y organizarla en formato “yyyy/MM/dd”, esta modificación se realizó en una nueva columna “FECHA\_HECHO\_C”. Las variables correspondientes tanto al código de departamento, código de municipio y cantidad de eventos se encuentran en formato número de esta manera no es necesario realizar ningún ajuste a estas variables.
* Para los valores registrados en Departamento y Municipio se seleccionó la opción UTF-8 a fin de que poder visualizar de manera correcta aquellos registros que cuenta con tilde.
* Se aplico un bloque Duplicate Row Filter donde no se evidencia ningún registro duplicado teniendo como resultado 69.137 registros.
* **Secuestro**
* Este archivo contiene información de la cantidad de eventos de secuestro registrados entre el año 1996 y 2022, discriminados por departamento y municipio cada uno de estos con su respectivo código de identificación.
* Dentro de las características de las variables se evidencia que la variable de fecha es de tipo string en formato dd/MM/yyyy hh:hh, para lo cual se realizó el respectivo ajuste al formato de fecha al “MM/dd/yyyy” mediante un bloque de String Manipulation para extraer solo la fecha y organizarla en formato “yyyy/MM/dd”, esta modificación se realizó en una nueva columna “FECHA\_HECHO\_C”. Las variables correspondientes tanto al código de departamento, código de municipio y cantidad de eventos se encuentran en formato número de esta manera no es necesario realizar ningún ajuste a estas variables.
* Para los valores registrados en Departamento y Municipio se seleccionó la opción UTF-8 a fin de que poder visualizar de manera correcta aquellos registros que cuenta con tilde.
* Se aplico un bloque Duplicate Row Filter donde no se evidencia ningún registro duplicado teniendo como resultado 18.642 registros.
* **Terrorismo**
* Este archivo contiene información de la cantidad de eventos de terrorismo registrados entre el año 2010 y 2022, discriminados por departamento y municipio, para este caso no se registran los códigos de departamento y municipio, se encuentra una nueva columna correspondiente al código DANE.
* Dentro de las características de las variables se evidencia que la variable de fecha es de tipo string en formato dd/MM/yyyy hh:hh, para lo cual se realizó el respectivo ajuste al formato de fecha al “MM/dd/yyyy” mediante un bloque de String Manipulation para extraer solo la fecha y organizarla en formato “yyyy/MM/dd”, esta modificación se realizó en una nueva columna “FECHA\_HECHO\_C”. Las variables correspondientes tanto al código de departamento, código de municipio y cantidad de eventos se encuentran en formato número de esta manera no es necesario realizar ningún ajuste a estas variables.
* Para los valores registrados en Departamento y Municipio se seleccionó la opción UTF-8 a fin de que poder visualizar de manera correcta aquellos registros que cuenta con tilde.
* Se aplico un bloque Duplicate Row Filter donde se evidencian 5 registros duplicados para un resultado 3.949 registros.
* **Homicidios**
* Este archivo contiene información de la cantidad de eventos de homicidios registrados entre el año 2010 y 2022, discriminados por departamento y municipio, para este caso no se registran los códigos de departamento y municipio, se encuentra una nueva columna correspondiente al código DANE.
* Dentro de las características de las variables se evidencia que la variable de fecha es de tipo string en formato d/MM/yyyy, se implemento un bloque String Manipulation donde se utilizo la función “padLeft($FECHA HECHO$,10,"0")” para agregar un cero a la izquierda a aquellos registro en el que el día solo contenía un digito y de esta manera dejar el mismo numero de caracteres para todos los registros de fechas. Luego se implemento un bloque String Manipulation para organizar los registros en el formato “yyyy/MM/dd”. Una vez organizada la información se implemento un bloque Stringe to Date&Time para cambiar el tipo de variable string a fecha.
* Para los valores registrados en Departamento y Municipio se seleccionó la opción UTF-8 a fin de que poder visualizar de manera correcta aquellos registros que cuenta con tilde.
* Se aplico un bloque Duplicate Row Filter donde se evidencian 18.583 registros duplicados para un resultado 240.740 registros.
* Se evidencia campos nulos y dos variables que son posibles unificar debido a que corresponden con el mismo significado (No reporta – No reportado)

****

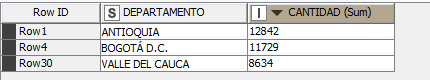
En la siguiente imagen se muestra el diagrama implementando con cada uno de los bloqueos para realizar el perfilamiento de cada uno de los cuatro archivos mediante el software Knime:

****

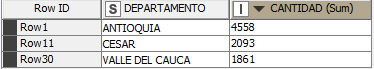
1. Responda las siguientes preguntas para cada uno de los conjuntos de datos seleccionados:

2.1. ¿Cuáles han sido los departamentos (TOP 3) más afectados en términos de cantidad de delitos cometidos en los últimos 5 años?

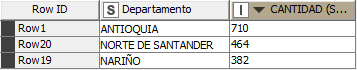
* **Extorsión**

****

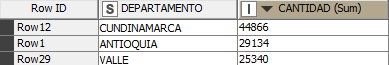
* **Secuestro**

****

* **Terrorismo**

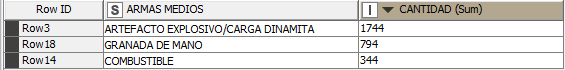
****

* **Homicidios**



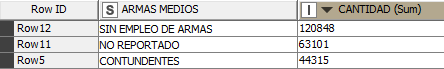
2.2. Para los casos en los que aplique, ¿cuál ha sido el arma o medio más común para cometer el delito?

* **Terrorismo**

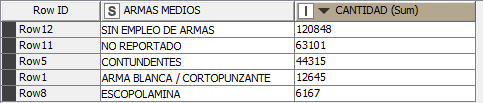


* **Homicidios**

Analizando el dataset de homicidios se tiene el siguiente top 3 de las armas, sin embargo, el primero lugar corresponde a sin empleo de armas y no reportado. Por esta razón se realizó un análisis un poco más a detalle de las armas asociadas a este delito:



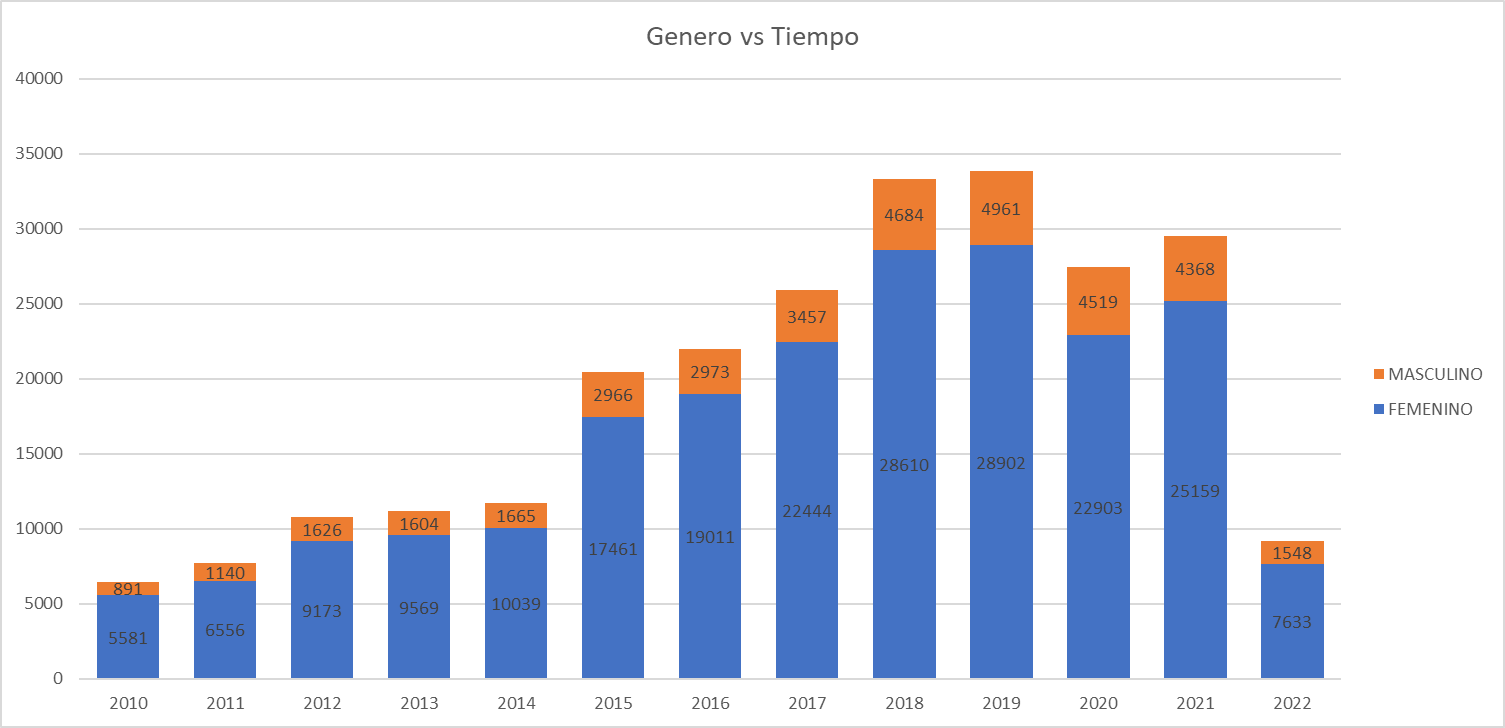
De esta manera podemos decir que el top 3 de armas utilizadas en los homicidios en primer lugar, son armas contundentes, en segundo armas blancas y en tercer lugar la escopolamina:



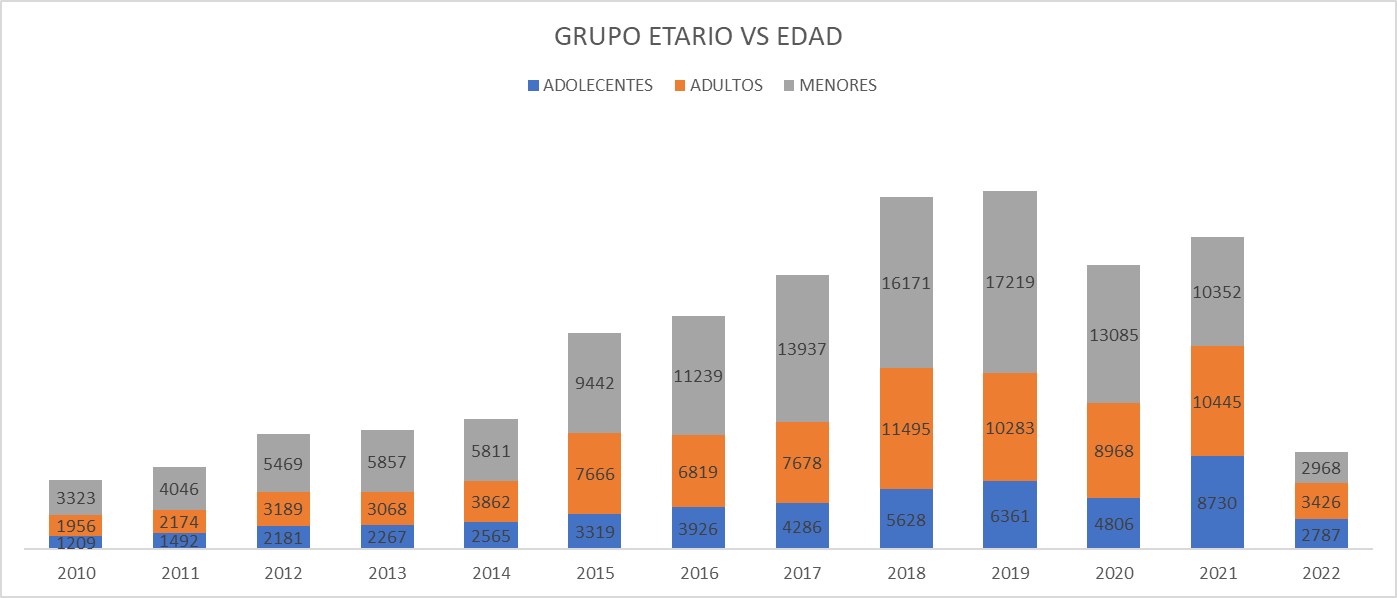
2.3. Para los casos en los que aplique, ¿cómo ha sido la proporción de géneros y grupos etarios que han estado involucrados en este tipo de delito? ¿Han variado con el paso de los años?

* **Homicidios**

En la siguiente grafica es posible identificar desde el año 2010 al 2022, la cantidad de casos que se han presentado por género. Se identifica de manera significativa que en todos los años se tiene un mayor porcentaje en el genero femenino. Adicionalmente es posible validar que entre los años 2018 y 2019 se presento el mayo numero de eventos de homicidio



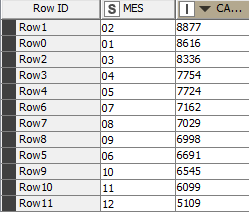
Respecto al análisis por grupo etario, es posible validar que la mayor parte de los homicidios a lo largo de los años a estado relacionada con los menores, seguido de los adultos y adolescentes.



2.4. ¿Se evidencia alguna tendencia para cometer dicho delito en algún mes particular del año?

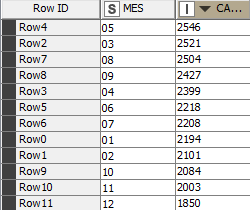
* Extorsión:

El mes donde se presenta un mayo número de casos de extorsión es el mes de febrero:



* Secuestro:

El mes que presenta un mayor numero de casos de secuestro es el mes de mayo:



* Terrorismo:

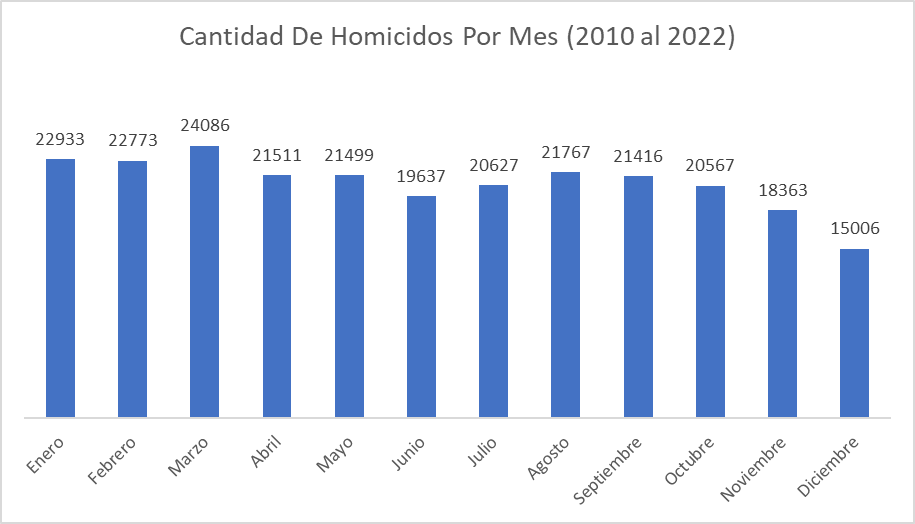
El mes que presenta un mayor número de casos de terrorismo es el mes de febrero:

Tabla

Descripción generada automáticamente

* Homicidio:

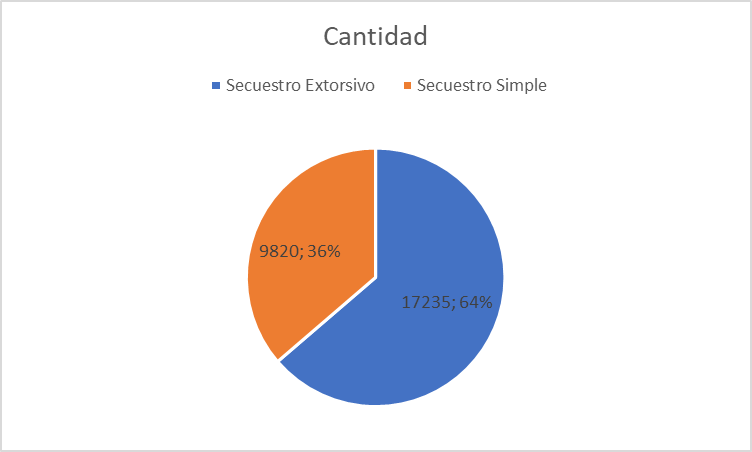
Se evidencia que la mayor cantidad de casos de homicidio se presenta en el mes de marzo con un total de 24.086 casos, seguido del mes de enero con 22.933:



2.5. Para los casos en los que se disponga del detalle del delito o de una descripción, como por ejemplo en delitos sexuales y secuestro, ¿cuáles son las descripciones o modalidades más comunes?

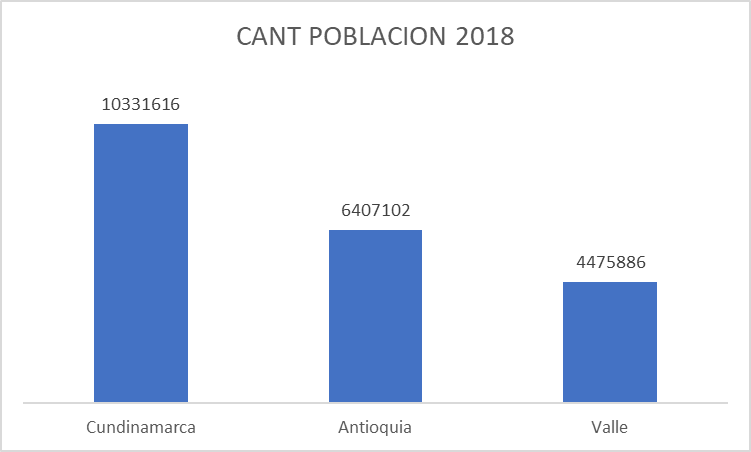
Para el caso de secuestro se identifica que el mayor tipo de delito corresponde al secuestro extorsivo correspondiente a un 64% de los casos presentados.

****

****

1. A partir de alguno de los conjuntos de datos seleccionados, visualice una serie de tiempo por año y mes que permita comparar la cantidad de delitos cometidos para los departamentos con mayor ocurrencia durante los últimos 5 años. Para que los resultados entre departamentos sean comparables, es importante que normalice las cantidades obtenidas por cantidad de habitantes. En este archivo puede encontrar la población por departamento para el año 2018. Asuma que la población no ha cambiado con el paso de los años.

Se selecciono el dataset de homicidios para realizar el análisis de este punto, la cantidad de población suministrada se extrajo de la página web del DANE, para los departamentos de Cundinamarca (Bogotá), Antioquia y Valle a fin de poder realizar la respectiva normalización:



Para realizar la normalización de los datos se dividió la cantidad de delitos sobre la cantidad de población para cada uno de los departamentos. En la siguiente grafica es posible validar después de realizar la normalización que existe un comportamiento muy similar entre los tres departamentos, adicionalmente se puede visualizar que no existe un comportamiento similar entre cada año, por el contrario, se tiene un leve incremento para los departamentos de Cundinamarca y Antioquia.

