

Prueba Analítica-Análisis de Datos Delictivos del Departamento de Antioquia

Fabián Yamith Tovar Altahona, CC. 1003952976

3 de Noviembre de 2024

Tratamiento previo de los datos

Antes de empezar con la solución de las preguntas propuestas, debó aclarar que los datos utilizados fueron organizados previamente en formato de tabla, con el uso de la herramienta Power Query. Esto se hizo con el fin de tener un único archivo .xlsx con la información de homicidios en Colombia a lo largo de todos los años solicitados (2016-2023), debido a que originalmente los datos de cada año se hallaban en diferentes documentos, y su tratamiento no era óptimo de esa manera. El archivo final lo llamé 'Homicidios_2016-2023.xlsx', el cual contiene 10 hojas dentro del mismo libro, de las cuales la más relevante recibe el nombre '2016-2023'. Con esta hoja creé el DataFrame con el que trabajé durante el proyecto.

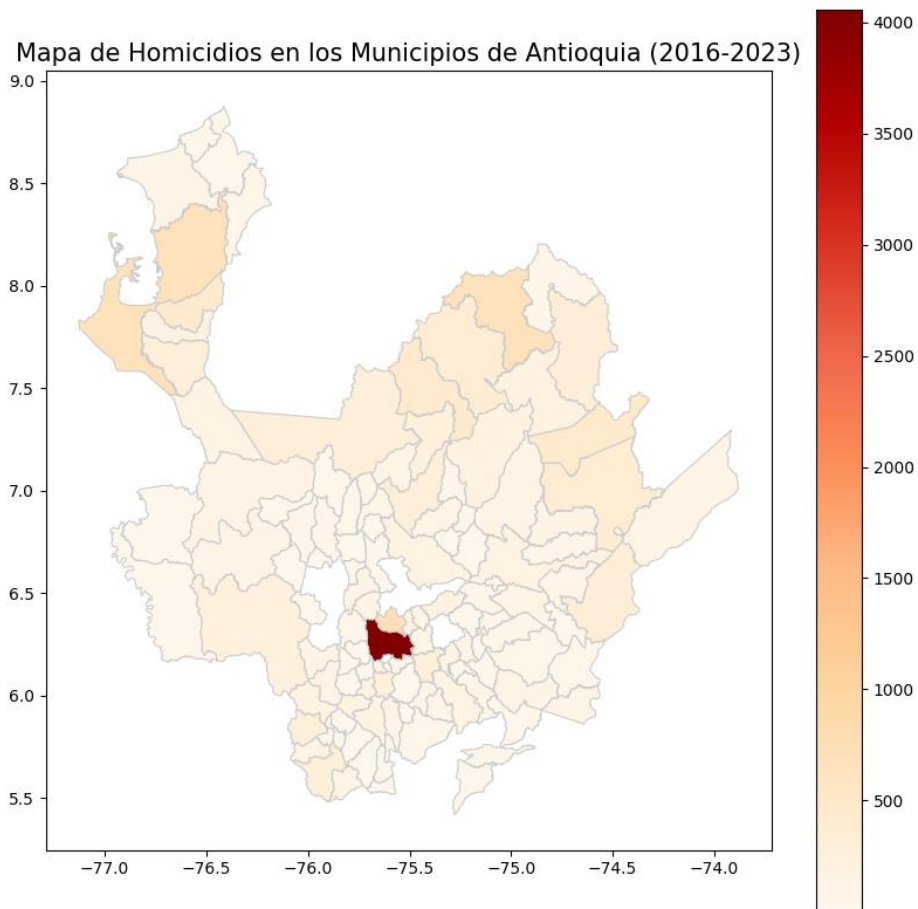
También accedí a la base de datos del DANE para adquirir una mínima información que usé sobre el marco Geoestadístico Nacional, para acceder al shapefile del nivel geográfico por municipios de Colombia (shapefile del año 2023). Y adicionalmente obtuve un documento de la base de datos de la Gobernación de Antioquia, en el que se aprecia la cantidad de habitantes por cada municipio del Departamento (Para el año 2020).

Los documentos mencionados anteriormente están anexados en la carpeta del proyecto (carpetas nombradas 'geoestadistico' y 'demografia-poblacion', ambas son imprescindibles para el correcto funcionamiento del Script .ipynb).

Resolución de Preguntas

1. *Observa la distribución de homicidios entre los municipios de Antioquia durante el período 2016-2023. ¿Qué patrones o anomalías detectas? ¿Qué preguntas de investigación se te ocurren a partir de estas observaciones? Describe y Resuelve cómo abordarías estas preguntas utilizando la información disponible.*

Para empezar, quise hacerme una idea visual de las variables y sus distribuciones, por lo que realicé un mapa de calor de los municipios de Antioquia sus cantidades totales de homicidios registrados entre 2016 y 2023. Así, pude obtener el siguiente gráfico

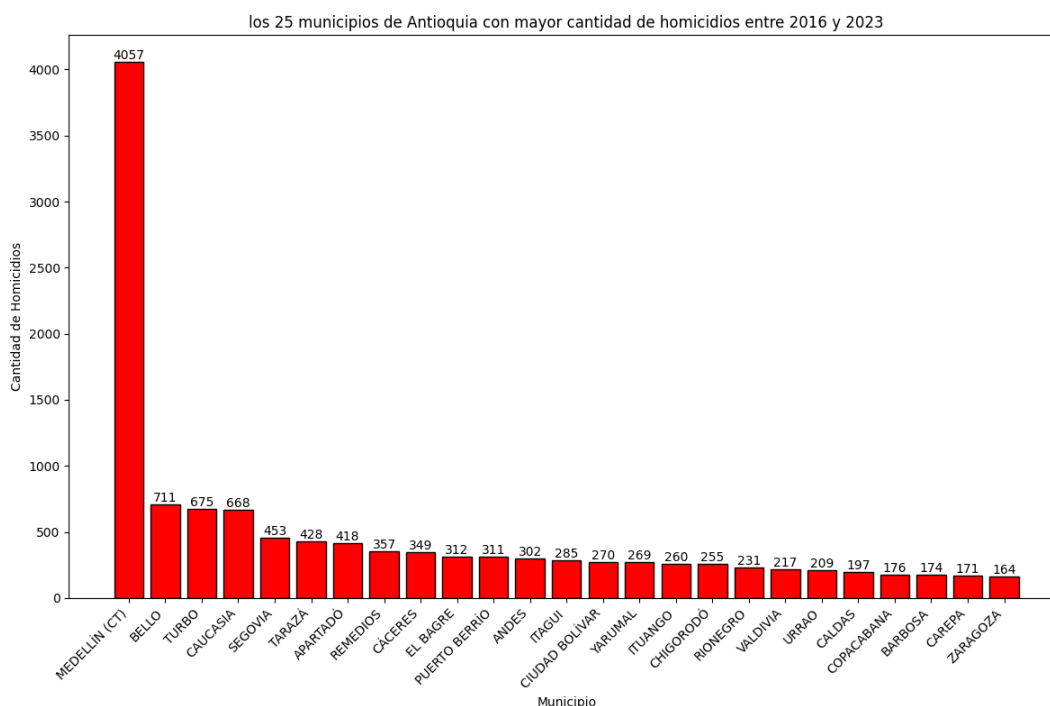


En el que se puede observar que la ciudad de Medellín supera ampliamente al resto de municipios en cuanto al número de casos de homicidios en el departamento de Antioquia, en el periodo de tiempo analizado. Dicho hecho no sorprende demasiado, porque según ‘infobae’ [1], Medellín fue la cuarta ciudad con mayor tasa de homicidio por cada 100 mil habitantes en el año 2022, con un valor de 4.6, sumado a que es la ciudad capital. Otro hecho que se visualiza es que las zonas

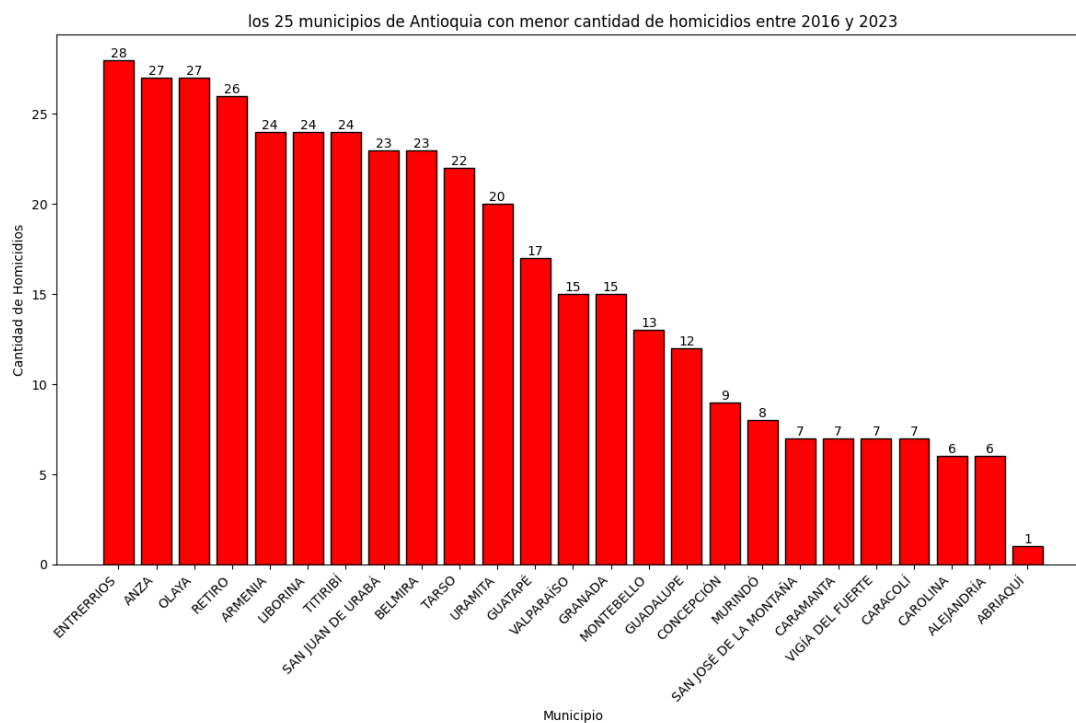
marginales del Departamento son las siguientes que más registran homicidios, por lo que me surge la primera pregunta

- **¿Por qué la mayoría de municipios que se ubican cerca de la frontera geográfica son los que más registran homicidios? (Evidentemente solo después del área metropolitana de Medellín).**

Como el esquema del mapa de calor no es suficientemente claro respecto a los números de homicidios y los municipios correspondientes, procedí a realizar una exploración mediante los gráficos de barras, tal como sigue



Este gráfico solo registra los primeros 25 municipios con más homicidios, ya que como Antioquia consta de 125 divisiones, la ilustración se saturaría y no sería clara. Luego, grafiqué inmediatamente la cola de la estadística, esto es, los 25 municipios con menor cantidad de homicidios.



En perspectiva el número de homicidios en la ciudad de Medellín parece totalmente desproporcionado. Pero aparentemente el tamaño de la población de los municipios es proporcional al número de homicidios, debido a que, de acuerdo a la base de datos de la gobernación de

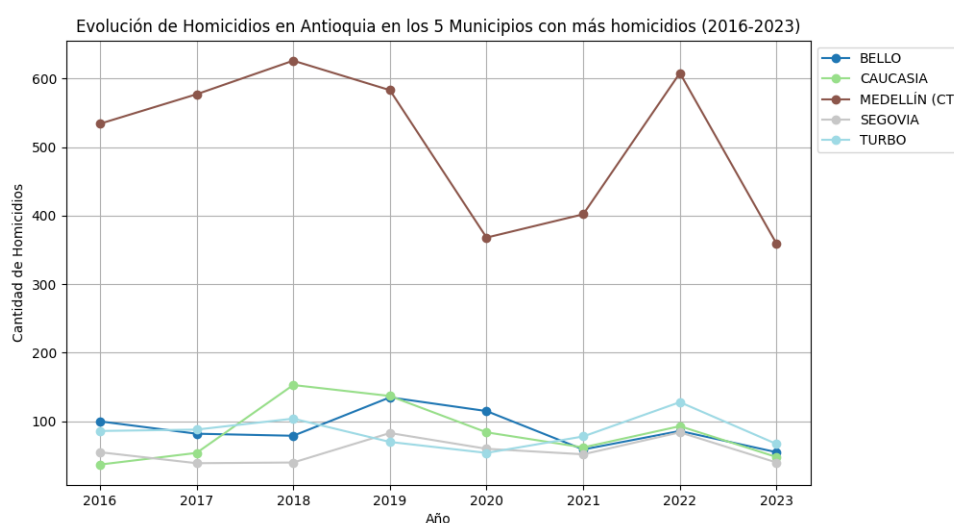
Antioquia, los municipios más poblados son los que aparecen primero en el diagrama de barras anterior. Pero ojo, es una apreciación que no ha sido comprobada; así pues, surge la pregunta

- **¿Existe una relación directamente proporcional o de algún tipo, entre la cantidad de homicidios y el número de habitantes del municipio?**

También quise analizar el histórico de homicidios por año y por municipio en el departamento, para responder otra pregunta:

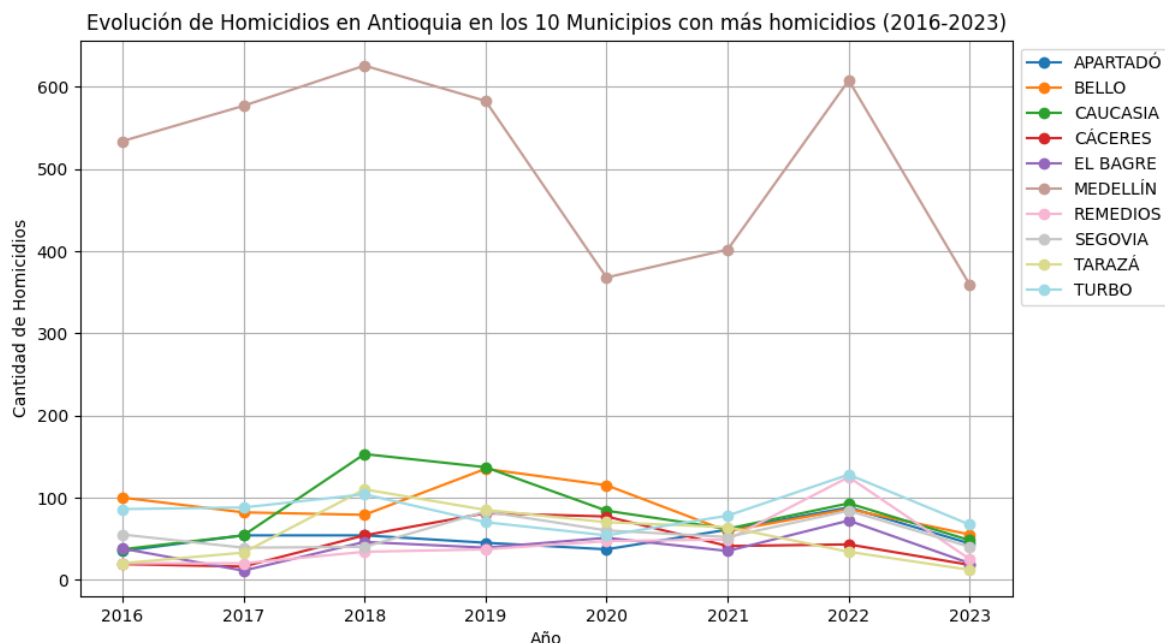
- **¿Cómo se han comportado los casos de homicidios en los últimos años en el departamento de Antioquia? ¿Han aumentado o Disminuido?**

Sin embargo, el hecho de que la cantidad de municipios sea tanta, dificulta un poco la tarea, por lo que nuevamente hice un gráfico para los primeros municipios con más homicidios, y su evolución en los años analizados.



Para los municipios con el mayor registro de homicidios entre 2016 y 2023 se observa una disminución particular entre los años 2020 y 2021 respecto al año anterior (por tal motivo, decimos que es un patrón interesante que pudiésemos examinar). Aparte, en el año 2023 se logró una disminución general de homicidios para dichos municipios. Si quisiéramos tazar el comportamiento

de una cantidad mayor de municipios para verificar esta tendencia, llegamos al siguiente gráfico: (que, aunque no tan legible, apoya mi observación)



La disminución de homicidios entre 2020 y 2021, así como en el año 2023, es notoria. Pero la tendencia de los años 2020 y 2021 podría tener explicación al visualizar la investigación publicada en la revista Nature [2] respecto al impacto del COVID-19 en diferentes países. Esto apunta a una disminución favorable en los índices de criminalidad y homicidios, en adición, un reporte del diario EL PAIS comprueba que ese impacto se sintió en las principales ciudades del estado colombiano en el periodo de cuarentena [3], que se dio precisamente entre los años 2020 y 2021. En cuanto al patrón del año 2023, requeriríamos más información relacionada con la implementación de controles judiciales, tal vez por parte de la policía nacional o del ejército, ya que puede ser un

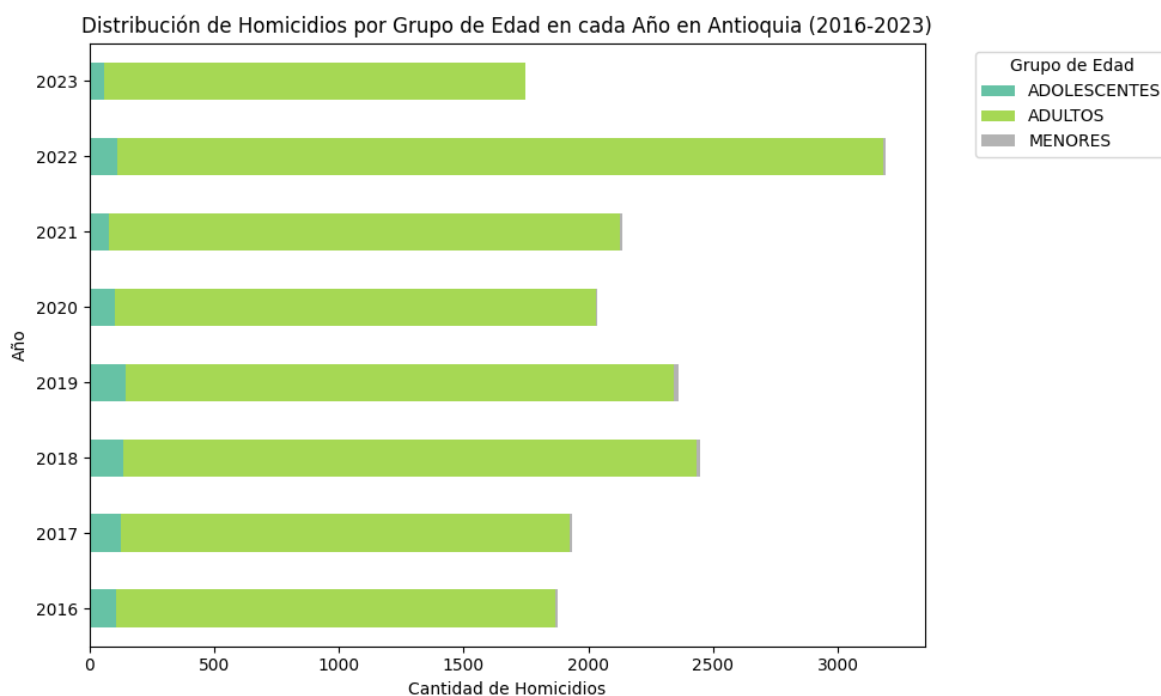
factore resaltante. Así como registros de turismo y migración; estos son ideas que se me ocurren para analizar tal fenómeno.

Una nueva pregunta que tuve estaba asociada a los grupos más afectados por este flagelo:

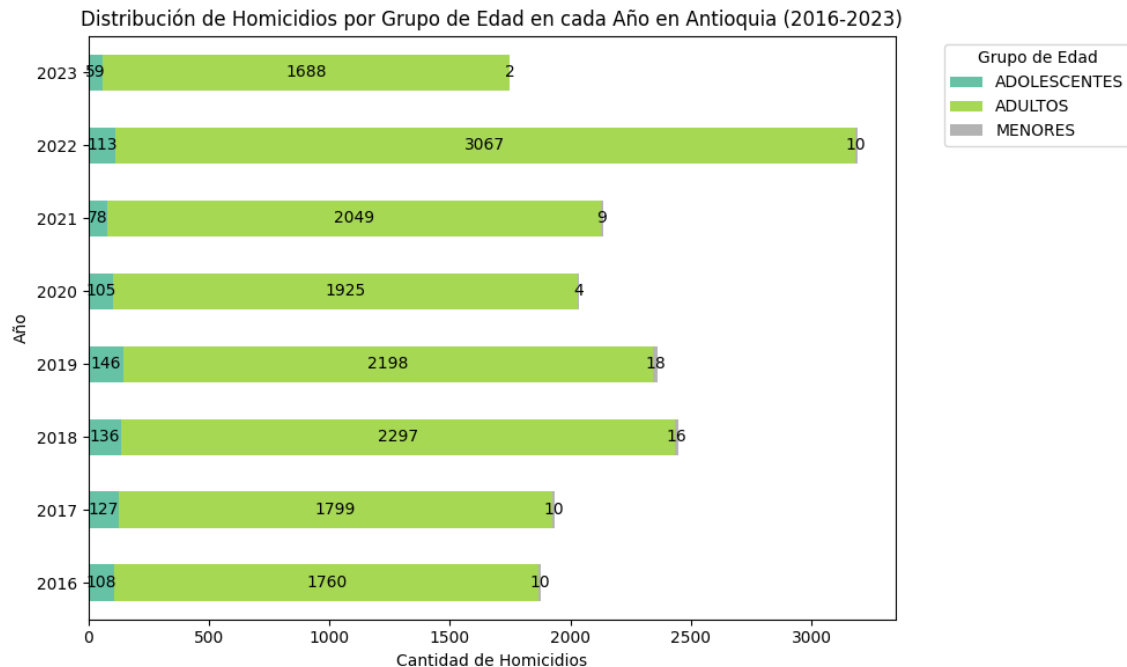
- **¿Cuáles son los grupos de personas sobre los cuáles más se registran homicidios?**

Y con el objetivo de abordar esta pregunta me dispuse a realizar distintos tipos de gráficos por categoría de datos presentes en la base de datos con la que estamos trabajando.

Primero pensé categorizar los grupos de persona por edad, por lo tanto, realicé un diagrama que me dio información de las edades de las personas sobre las que se han registrado más homicidios a lo largo del intervalo de tiempo que tenemos.

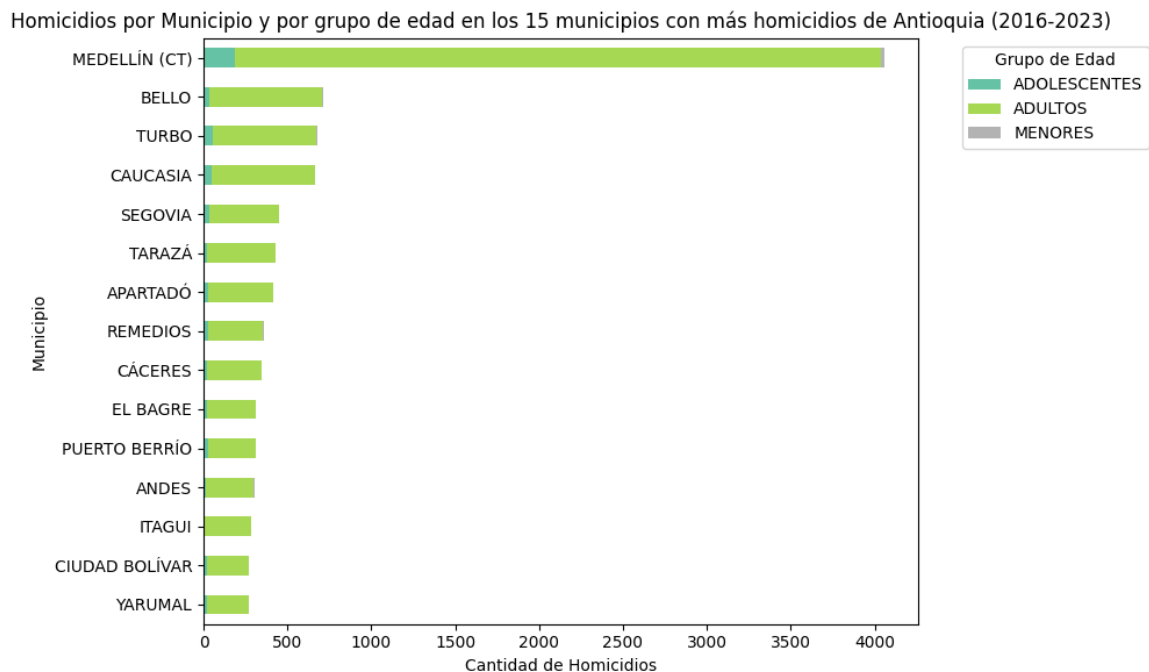


Pero para visualizar un poco mejor las cifras, podemos ayudarnos con este otro gráfico

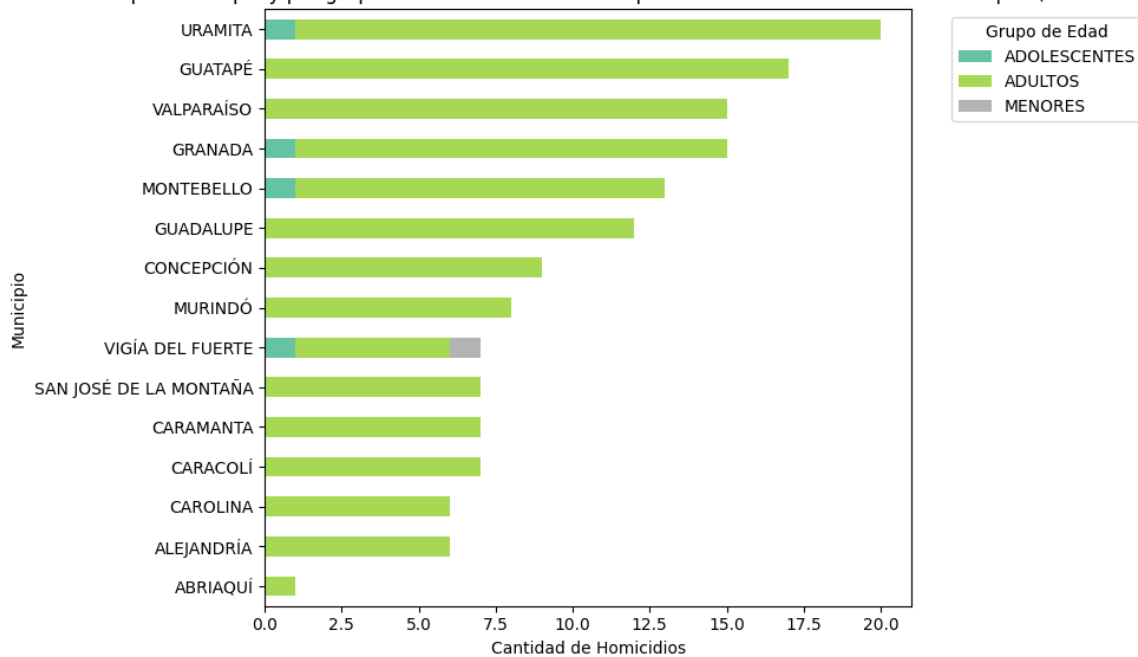


Es evidente que, por este lado, la pregunta ya quedó resuelta. El grupo de edad más vulnerable a los homicidios son los adultos, representando en cada año la gran mayoría.

Y si quisiéramos revisar este mismo punto, pero esta vez no categorizado por año, sino por municipio, el resultado apoyaría más lo evidente de la observación pasada, de la siguiente manera:



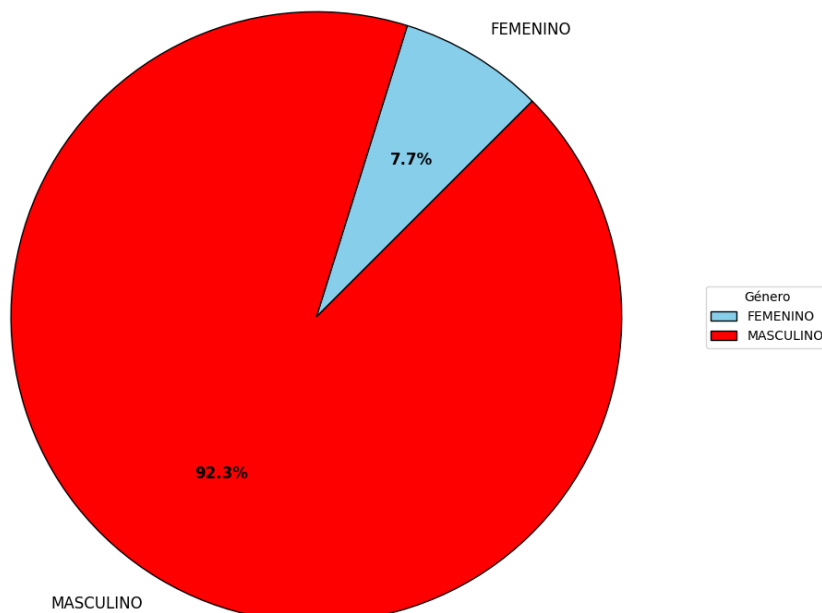
Homicidios por Municipio y por grupo de edad en los 15 municipios con menos homicidios de Antioquia (2016-2023)



Note que nuevamente los adultos son aquellos en los que más homicidios se registran. Incluso en aquellos municipios donde menos han ocurrido homicidios.

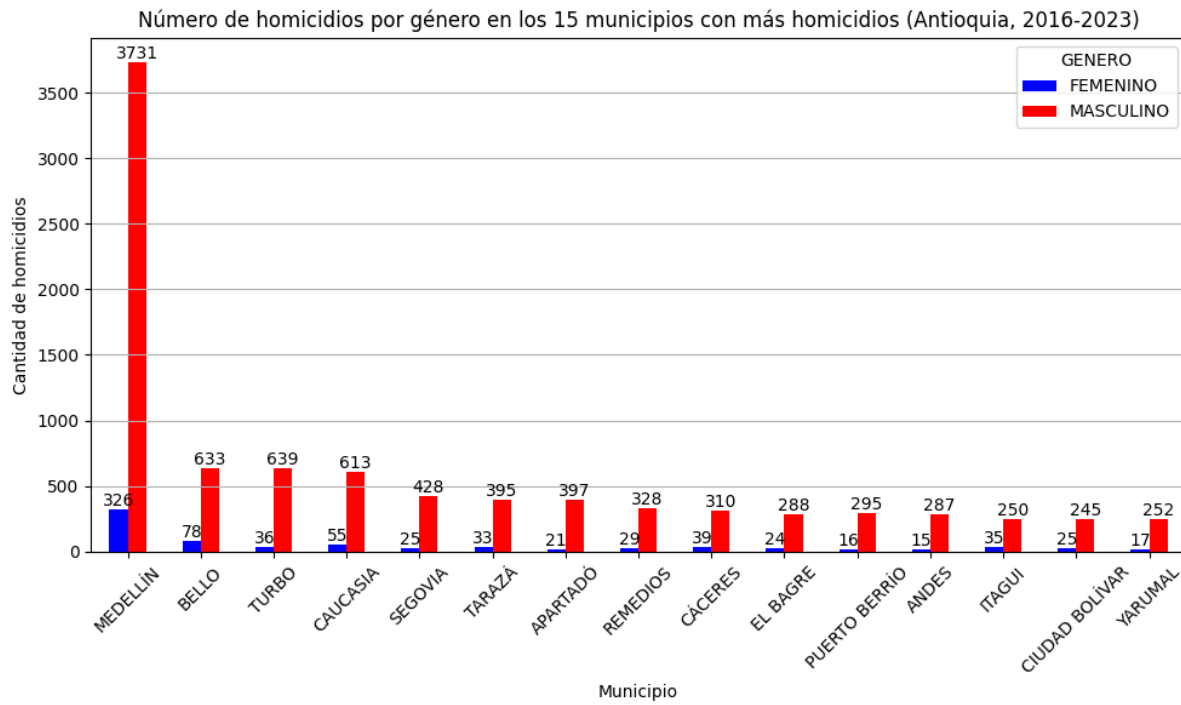
Ahora, siguiendo con la respuesta de la pregunta, me propuse a analizar las estadísticas, pero esta vez por género. Para identificar cuál es el género más común de las personas sobre quienes se comete homicidio, realicé un gráfico de pastel que muestra los porcentajes de forma clara

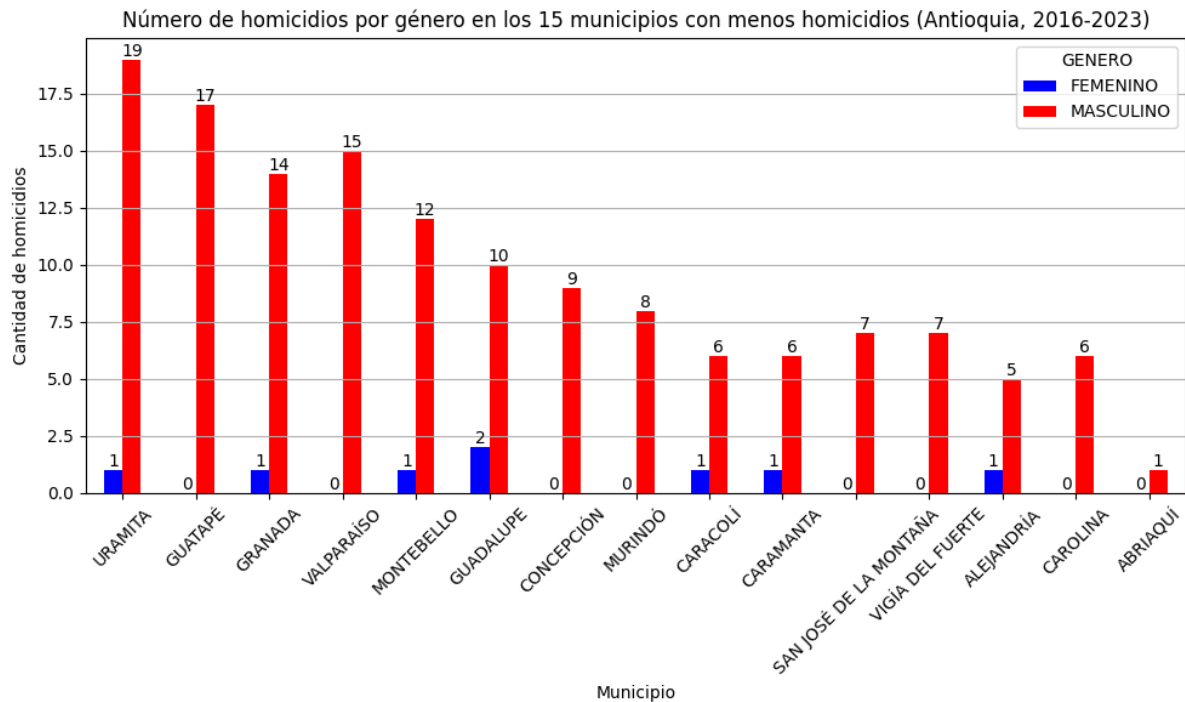
Distribución de Homicidios por Género en Antioquia (2016-2023)



Y la respuesta es nuevamente clara, más del 92% de los homicidios que ocurrieron entre los años 2016 y 2023 en el departamento de Antioquia fueron sobre hombres, así que, por género, el grupo sobre el cual más homicidios suceden es

en los hombres. Y el comportamiento de este flagelo se evidencia en todos los municipios del departamento, como se ve en los gráficos a continuación:

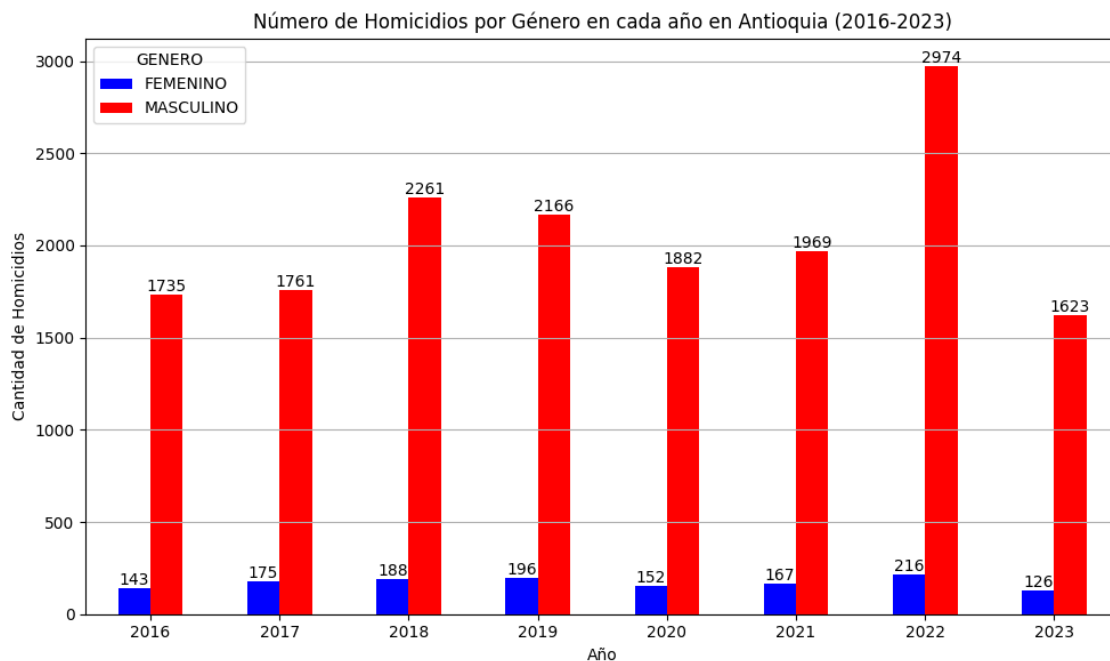




Con base en estos gráficos es fácil ver que incluso en los lugares donde hubo un solo homicidio, el afectado fue un hombre. Luego, nace otra pregunta que vale a pena investigar, al ser tan resaltante

- **¿Cuáles son los factores que hacen que los hombres sean las principales víctimas de homicidios en el departamento de Antioquia a lo largo de los años?**

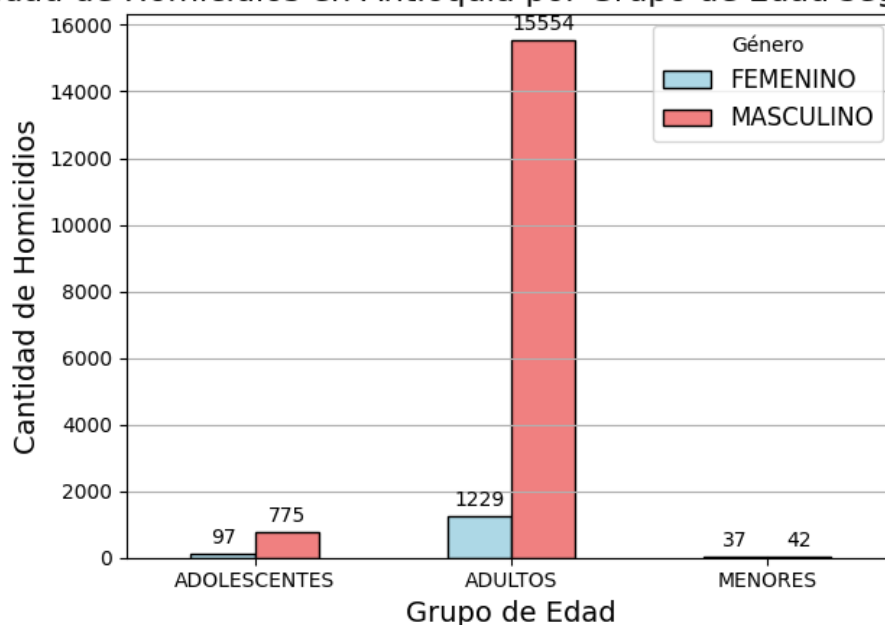
Y digo a lo largo de los años porque también se puede comprobar gráficamente que a lo largo de los años (2016-2023) los hombres han sido los principales afectados, tal como se aprecia en la imagen que sigue:



Para responder esta pregunta yo apelaría a buscar información sobre el número de población en el departamento por género, y para aclarar además por qué los adultos son las principales víctimas, también indagaría sobre la cantidad de habitantes en el departamento agrupada por edad. Tal vez con eso podríamos llegar a una relación entre el número de homicidios a cierto grupo de personas y la cantidad de habitantes de dicho grupo en la región. Esto sería una particularidad de la respuesta a la segunda pregunta, la cual no he respondido, pero lo haré en breve.

Con base en los datos que estamos analizando, es un hecho que entonces, los grupos de personas sobre los cuales se registran más homicidios son los hombres y adultos, y siendo un poco más específicos, en los hombres adultos. El diagrama de cantidad de homicidios por género según el grupo de edad nos da una idea más clara de esto.

Cantidad de Homicidios en Antioquia por Grupo de Edad según el Género

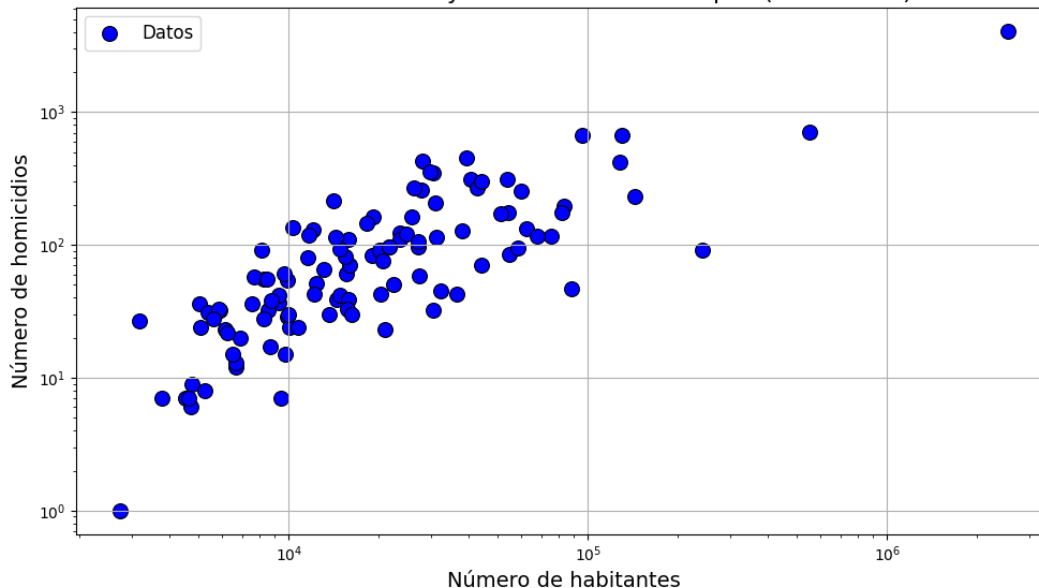


Aquí visualizamos que hasta en los adolescentes y menores, quienes más sufren homicidios son los hombres. **¿Qué medidas de precaución se deberían tomar sobre esta población? Y ¿De qué forma se pueden implementar planes de acción para reducir dicha estadística?**

Por último, para finalizar con la primera sección y responder la pregunta que hacía falta, sobre la relación entre los casos de homicidios y el número de habitantes. Tomé el registro del número de habitantes en cada municipio de Antioquia, desde la base de datos de libre acceso de la gobernación,

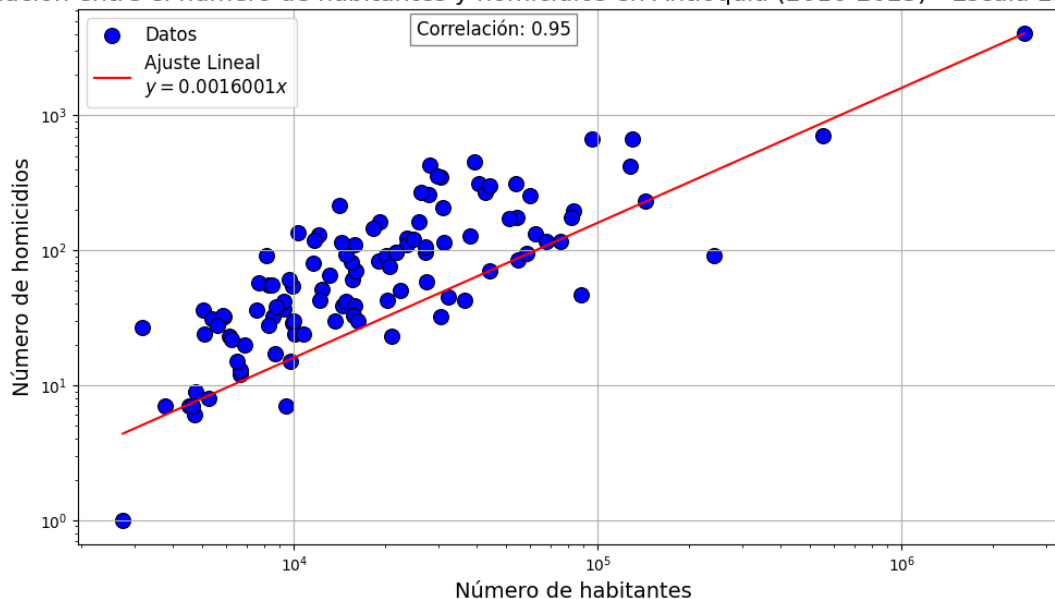
y me di a la tarea de relacionar los municipios con su respectivo número de habitantes para ver su comportamiento respecto al número total de homicidios, en un gráfico de dispersión.

Relación entre el número de habitantes y homicidios en Antioquia (2016-2023) - Escala Logarítmica



A simple vista parece que, si tienen una relación estas variables, por lo que procedí a hacer una regresión lineal a fin de conocer qué tanto de ajustan estos datos, o qué tanto se relacionan los homicidios con el número de habitantes. Entonces, usando una de las herramientas que me brinda la librería de Python 'scipy' obtuve el siguiente ajuste:

Relación entre el número de habitantes y homicidios en Antioquia (2016-2023) - Escala Logarítmica



Luego, se prueba que la cantidad de homicidios está relacionada linealmente con el número de habitantes, pero con un coeficiente de correlación de Pearson de 0.95, es decir, la gran mayoría de los datos siguen esta tendencia. Este mismo resultado fue el obtenido en un estudio hecho por Luis David Ramirez sobre la tasa de homicidios en México [4].

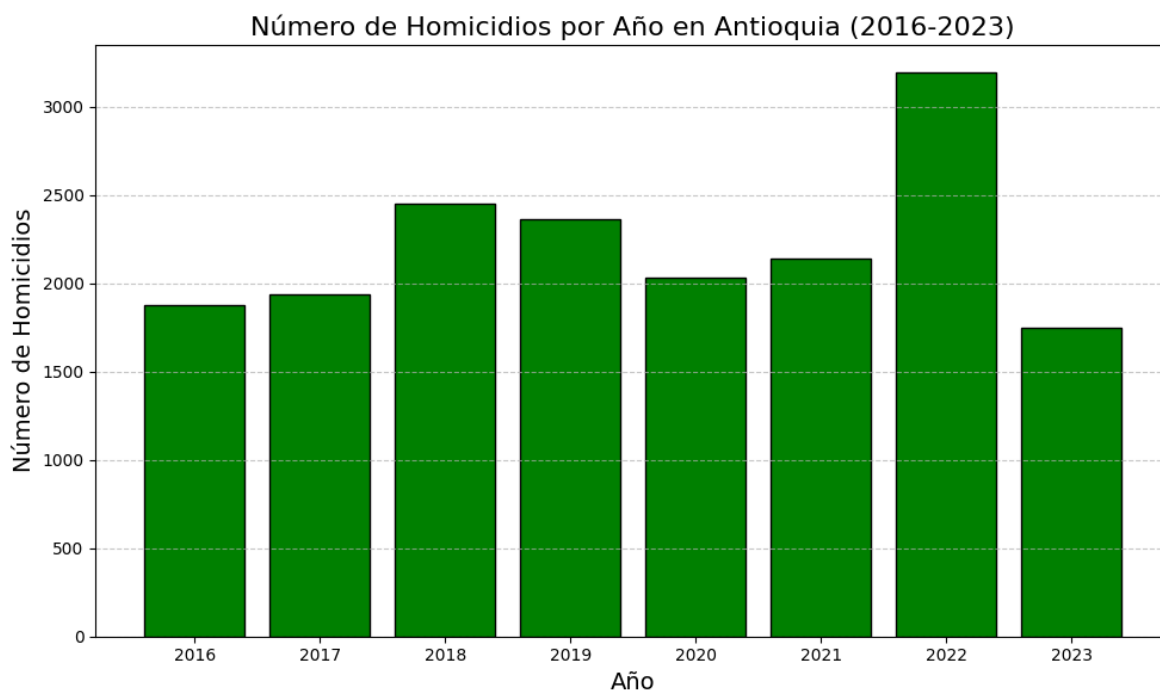
2. *Considera la base de datos de homicidios y reflexiona sobre la posible presencia de datos atípicos. ¿Qué características podrían indicar la presencia de estos datos? ¿Cómo podrían afectar los datos atípicos a la validez de las conclusiones obtenidas? Adicionalmente, imagina un escenario donde faltan los datos de homicidios para los meses de septiembre y octubre en 2023. ¿Qué técnicas utilizarías para rellenar estos datos faltantes? Justifica tu elección y discute las posibles limitaciones de cada técnica.*

En esta base de datos se podrían manifestar datos atípicos de la siguiente manera:

Podríamos buscar municipios cuyas tasas de homicidios sean muy grandes o muy pequeñas con relación a los demás, por ejemplo, Medellín tiene una tasa de Homicidios extremadamente elevada con relación los otros municipios, y aunque sea la ciudad capital, es interesante indagar sobre las razones de este acontecimiento. Una idea que se me viene a la mente es realizar un gráfico de ‘cajas y bigotes’ entre la cantidad de homicidios por año por y cada municipio, para determinar si los rangos de homicidios por año en la ciudad de Medellín contienen algún valor fuera del rango

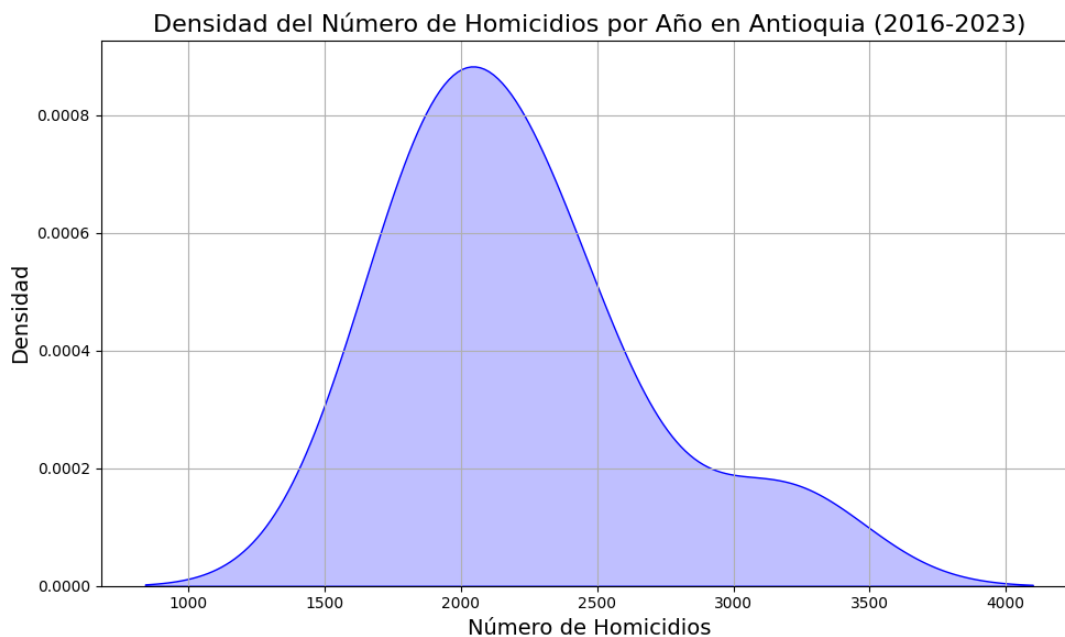
intercuartil o de los ‘bigotes’ del diagrama. De esta forma podríamos analizar si es buena idea descartar los datos atípicos que se presenten.

Otro indicio de la posible presencia de datos atípicos son los cambios precipitados en la tendencia de los homicidios. Un caso que sí se puede ejemplificar en esta sección es la distribución de los homicidios en los años, esta tiene la siguiente estructura:



Nos podemos percatar aquí de que desde el año 2018 los homicidios tenían una tendencia a disminuir, incluso en 2021 subieron pero muy poco en comparación a lo que bajaron entre 2019 y 2020. No obstante, en el año 2022 hay un aumento inesperado que cortó dicha tendencia, porque en ese año los homicidios superaron la barrera de los 2500, y tomó valores superiores a 3000 y más cercanos a 3500. Por lo tanto, si realizamos un gráfico de la densidad de homicidios por año,

obtendremos una distribución que no será totalmente simétrica, esto podría representar un valor atípico.



Y como lo predijimos, la distribución del número de homicidios no es totalmente simétrica, y es que hay una ‘anomalía’ precisamente entre los valores de 3000 y 3500, que concuerda con el gráfico de barras en el año 2022, donde se llegó a más o menos a ese rango de homicidios. Así, esto no es una prueba de que en estos años los homicidios alcanzan un valor atípico, pero sí es una alarma de que podría estar ocurriendo algo anormal.

Es una buena práctica deshacerse de los datos atípicos, con el fin de hacer un análisis estadístico no distorsionado. Sobre todo pueden influir mucho en las medidas estadísticas como promedios, o desviaciones estándar, en este caso el posible dato atípico podría afectar el promedio de homicidios por año. Asimismo, esto puede generar problemas en caso de que hallan datos faltantes, debido a que se podrían originar predicciones erróneas a la hora de rellenar información incompleta.

En cuanto a el tratamiento de datos faltantes (entre septiembre y octubre de 2023), existen varias formas de tratar esto, y también depende del tipo de datos que esté faltando. Por ejemplo, en este

caso los datos faltantes son los homicidios entre los meses establecidos, entonces algunas de las técnicas que se suelen utilizar para tipos de datos numéricos como este son:

- Insertar la media o la mediana:
Consiste en rellenar los datos faltantes con el valor promedio o la mediana de los meses anteriores. Esta es una técnica rápida y tampoco afecta la distribución de los datos en general. Sin embargo, esta puede que no tenga muy en cuenta podrían variar los datos.
- Insertar valores anteriores:
Se trata de rellenar los datos faltantes utilizando los valores del mes anterior, esto es, del mes de agosto del año 2023. Lo bueno de esta técnica es que la tendencia más reciente. El problema es que puede asumir que la tendencia es lineal o que no hay variación, lo cual podría ser incorrecto.
- Interpolación lineal:
Consiste en usar una interpolación lineal para añadir los datos de los homicidios faltantes, con base en los datos siguientes y anteriores. Esto es útil si los datos tienen una tendencia lineal, pero, de no ser así, estamos asumiendo que hay una relación lineal entre los meses, lo cual no es necesariamente cierto, porque un mes podría tener variaciones inesperadas.
- Modelos predictivos (regresión o modelos ARIMA):
También podríamos usar modelos estadísticos como la regresión o modelos como (ARIMA) para predecir los valores de homicidios en los meses faltantes. Lo bueno es que se tienen en cuenta las tendencias históricas y los patrones estacionales. La desventaja es que requieren mayor cantidad de conocimientos previos, y los resultados dependen de la cantidad de datos disponibles. Este es un método que se suele usar en Machine Learning. (Al igual que la técnica KNN, la cuál utilizo como ejemplo al final del Script que anexo con este informe).

3. *Investiga si existe una relación entre el tipo de arma utilizada en homicidios y el género de la víctima en Medellín. Formula una hipótesis sobre esta relación y realiza una prueba estadística apropiada para determinar si hay evidencia suficiente para respaldarla. Discute los resultados de la prueba y las posibles implicaciones de la relación (o la falta de ella)*

entre estas dos variables categóricas. ¿Qué conclusiones puedes extraer sobre los patrones de homicidios en Medellín a partir de este análisis?

Hipótesis nula: No existe relación significativa entre el tipo de arma utilizada en los homicidios y el género de la víctima en Medellín.

Hipótesis alternativa: Existe una relación significativa entre el tipo de arma utilizada en los homicidios y el género de la víctima en Medellín.

Ya que ambas variables son categóricas (ARMA MEDIO y GÉNERO), la prueba estadística más apropiada es la prueba de Chi-cuadrado de independencia. Esta prueba nos permite evaluar si las

distribuciones de las frecuencias observadas en estas dos variables categóricas son independientes entre sí o si están asociadas.

Para esto primero filtré los datos a solo aquellos que pertenecieran a Medellín. Y luego realicé una tabla de contingencia para realizar la prueba de ajuste Chi-cuadrado con base en una de las librerías de python que me permite hacer esto de una manera más rápida.

Tabla de contingencia - Arma empleada - Género		
ARMA MEDIO	FEMENINO	MASCULINO
ALMOHADA	2	0
ARMA BLANCA / CORTOPUNZANTE	126	893
ARMA DE FUEGO	110	1703
ARTEFACTO EXPLOSIVO/CARGA DINAMITA	0	1
BOLSA PLASTICA	2	21
CINTAS/CINTURON	4	7
CONTUNDENTES	68	335
CUERDA/SOGA/CADENA	6	15
GASES	1	0
GRANADA DE MANO	0	1
PRENDAS DE VESTIR	0	1
SIN EMPLEO DE ARMAS	2	4
SUSTANCIAS TOXICAS	0	1
VENENO	0	1

Finalmente obtuve lo siguiente:

```

Estadístico Chi-cuadrado: 108.88
Valor p: 3.093102674405681e-17
Grados de libertad: 13
Tabla esperada:
[[1.94309927e-01 1.80569007e+00]
 [9.90009080e+01 9.19999092e+02]
 [1.76141949e+02 1.63685805e+03]
 [9.71549637e-02 9.02845036e-01]
 [2.23456416e+00 2.07654358e+01]
 [1.06870460e+00 9.93129540e+00]
 [3.91534504e+01 3.63846550e+02]
 [2.04025424e+00 1.89597458e+01]
 [9.71549637e-02 9.02845036e-01]
 [9.71549637e-02 9.02845036e-01]
 [9.71549637e-02 9.02845036e-01]
 [5.82929782e-01 5.41707022e+00]
 [9.71549637e-02 9.02845036e-01]
 [9.71549637e-02 9.02845036e-01]]

```

Ya que el valor p es menor que 0.05 la hipótesis principal se rechaza, y por lo tanto demostramos que sí existe una relación entre el tipo de arma del homicidio y el género.

Referencias

1. <https://www.infobae.com/america/colombia/2022/06/09/estas-son-las-ciudades-con-las-tasas-mas-altas-de-homicidios-en-colombia/>
2. E. Nivette, R. Zahnow, R. Aguilar, A. Ahven, S. Amram, B. Ariel, M. J. A. Burbano, R. Astolfi, D. Baier, H.-M. Bark, J. E. H. Beijers, M. Bergman, G. Breetzke, I. A. Concha-Eastman, S. Curtis-Ham, R. Davenport, C. Díaz, D. Fleitas, M. Gerell, K.-H. Jang, J. Kääriäinen, T. Lappi-Seppälä, W.-S. Lim, R. L. Revilla, L. Mazerolle, G. Meško, N. Pereda, M. F. T. Peres, R. Poblete-Cazenave, S. Rose, R. Svensson, N. Trajtenberg, T. van der Lippe, J. Veldkamp, C. J. V. Perdomo y M. P. Eisner, "A global analysis of the impact of COVID-19 stay-at-home restrictions on crime", *Nat. Hum. Behav.* **5**(7), 868–877 (2021). <https://doi.org/10.1038/s41562-021-01139-z>
3. <https://elpais.com/ciencia/2021-06-02/los-confinamientos-por-la-covid-19-redujeron-la-criminalidad-en-un-37.html>
4. Crimen y economía: análisis de la tasa de homicidio en México a partir de variables económicas (2000, 2005, 2010). Author(s): Luis David Ramírez de Garay. Source: *Estudios Sociológicos*, Vol. 32, No. 96 (septiembre-diciembre, 2014), pp. 505-540. Published by: El

Colegio de Mexico. Stable URL: <http://www.jstor.org/stable/24368226>. Accessed: 17-07-2017 19:20 UTC