

COLOMBIANOS DETENIDOS EN EL EXTERIOR.

1. Definición del Problema

El objetivo de este proyecto es analizar la situación de los colombianos detenidos en el exterior. Se busca comprender las tendencias y patrones en las detenciones, considerando aspectos como la evolución en el tiempo, la distribución geográfica y el perfil de los detenidos.

1.1. Importancia del estudio:

- Identificación de tendencias: Entender cómo varían las detenciones a lo largo del tiempo puede revelar cambios en las políticas internacionales y en la aplicación de la ley.
- Enfoque geográfico: Conocer en qué regiones o países se concentran las detenciones permite orientar mejor los esfuerzos de asistencia consular y diplomática.
- Perfil de los detenidos: Analizar características demográficas y otras variables relevantes ayuda a diseñar programas de apoyo y estrategias de intervención para proteger los derechos de los ciudadanos.
- Perfil y clasificación de delitos: Analizar la clase de delitos que motivan o están asociados a estas detenciones es crucial para identificar patrones delictivos. Esta información es valiosa para diseñar campañas preventivas que aborden específicamente las causas y contextos en que se cometen ciertos delitos.

1.2. ¿Por qué vale la pena investigarlo?

- Impacto en la política exterior y relaciones internacionales: Entender dónde y bajo qué circunstancias se encuentran detenidos colombianos en el exterior puede aportar a la formulación de políticas de protección y asistencia consular.
- Análisis de tendencias y patrones: El estudio de estos datos puede revelar tendencias en la detención a lo largo del tiempo, identificar países con mayores registros y tipos de delito más frecuentes.
- Apoyo a la toma de decisiones: Los hallazgos podrían contribuir a la optimización de estrategias de intervención y cooperación con organismos internacionales, así como a mejorar el apoyo consular a los detenidos.

1.3. Objetivos y Preguntas de Investigación

Objetivos del estudio:

- Analizar la evolución de las detenciones de colombianos en el exterior a lo largo del tiempo.
- Identificar los países y regiones con mayor incidencia de detenciones.
- Evaluar el perfil demográfico de los detenidos, considerando variables como género y grupo de edad.
- Determinar la tipología de delitos asociados a estas detenciones, con el fin de orientar campañas preventivas y estrategias de intervención.

Preguntas de Investigación:

- ¿Cómo varía el número de detenidos a lo largo del tiempo y existen picos o tendencias claras?
- ¿Cuáles son los países o regiones donde se concentra la mayor cantidad de detenciones?
- ¿Qué perfil demográfico (por ejemplo, género y grupo de edad) se observa en los detenidos?
- ¿Qué tipos de delitos son más frecuentes en el registro de detenciones y cómo se relacionan con la ubicación geográfica?
- ¿Qué implicaciones pueden derivarse de estos patrones para el diseño de políticas de asistencia y campañas preventivas?

2. Contribución del Análisis de Datos

El análisis de datos es una herramienta clave para comprender y abordar la problemática de los colombianos detenidos en el exterior. A través de técnicas estadísticas y visualizaciones, se puede transformar un gran volumen de registros en información significativa que permita tomar decisiones informadas. En este proyecto, el análisis de datos contribuirá de las siguientes maneras:

- Detección de Tendencias Temporales: Al examinar la evolución de las detenciones a lo largo del tiempo, se podrán identificar períodos con aumentos o disminuciones notables. Esto permitirá relacionar cambios en las cifras de detenciones con eventos internacionales o modificaciones en las políticas de extradición y repatriación.
- Análisis Geográfico: Mediante la representación de datos georreferenciados (por ejemplo, usando mapas de calor o gráficos de dispersión), se podrán identificar los países o regiones donde se concentran las detenciones. Este enfoque geográfico es fundamental para orientar los esfuerzos consulares y diplomáticos hacia las áreas más críticas.
- Segmentación Demográfica y por Tipo de Delito: El análisis de variables como género, grupo de edad y la clasificación de delitos ayudará a perfilar a los detenidos. Esto no solo permite comprender mejor a la población afectada, sino que también facilita la creación de campañas preventivas y programas de asistencia adaptados a las características específicas de cada grupo.
- Identificación de Relaciones y Patrones Ocultos: Aplicando técnicas estadísticas, es posible identificar correlaciones entre diferentes variables (por ejemplo, entre el tipo de delito y la ubicación geográfica). Estas relaciones pueden revelar factores comunes o causales que, de ser abordados, podrían contribuir a reducir la incidencia de detenciones.

2.1. Ideas que se esperan obtener del análisis:

- Un panorama claro de cómo varían las detenciones a lo largo del tiempo, permitiendo detectar periodos críticos.
- Mapas y gráficos que muestren la distribución geográfica de los detenidos, identificando zonas con alta incidencia.
- Un perfil demográfico detallado de la población detenida, que incluya información sobre género y grupo de edad.
- Un análisis de los tipos de delitos más frecuentes, para orientar campañas preventivas específicas.
- Recomendaciones basadas en evidencia que apoyen la formulación de políticas de protección y estrategias de asistencia consular.

```
In [44]: #verificando La ubikcación.
import os
print(os.getcwd())
```

C:\Users\steph

```
In [46]: #cargando La base (tuve algunos inconvenientes)
import pandas as pd
import numpy as np

# Ruta
ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'

# Cargar el CSV
df = pd.read_csv(ruta, sep=',', encoding='utf-8')

# Mostrar Las primeras filas para verificar
df.head()
```

```
-----
ParserError                                Traceback (most recent call last)
Cell In[46], line 9
      6 ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'
      8 # Cargar el CSV
----> 9 df = pd.read_csv(ruta, sep=',', encoding='utf-8')
      11 # Mostrar las primeras filas para verificar
      12 df.head()

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1026, in read_csv(filepath_or_buffer, sep, delimiter, header, names, index_col, usecols, dtype, engine, converters, true_values, false_values, skipinitialspace, skiprows, skipfooter, nrows, na_values, keep_default_na, na_filter, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_col, date_parser, date_format, dayfirst, cache_dates, iterator, chunksize, compression, thousands, decimal, lineterminator, quotechar, quoting, doublequote, escapechar, comment, encoding, encoding_errors, dialect, on_bad_lines, delim_whitespace, low_memory, memory_map, float_precision, storage_options, dtype_backend)
    1013 kwds_defaults = _refine_defaults_read(
    1014     dialect,
    1015     delimiter,
    (...)
    1022     dtype_backend=dtype_backend,
    1023 )
    1024 kwds.update(kwds_defaults)
-> 1026 return _read(filepath_or_buffer, kwds)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:626, in _read(filepath_or_buffer, kwds)
    623 return parser
    625 with parser:
--> 626     return parser.read(nrows)

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\readers.py:1923, in TextFileReader.read(self, nrows)
    1916 nrows = validate_integer("nrows", nrows)
    1917 try:
    1918     # error: "ParserBase" has no attribute "read"
    1919     (
    1920         index,
    1921         columns,
    1922         col_dict,
-> 1923     ) = self._engine.read( # type: ignore[attr-defined]
    1924         nrows
    1925     )
    1926 except Exception:
    1927     self.close()

File ~\anaconda3\Lib\site-packages\pandas\io\parsers\c_parser_wrapper.py:234, in CParserWrapper.read(self, nrows)
    232 try:
    233     if self.low_memory:
--> 234         chunks = self._reader.read_low_memory(nrows)
    235         # destructive to chunks
    236         data = _concatenate_chunks(chunks)

File parsers.pyx:838, in pandas._libs.parsers.TextReader.read_low_memory()

File parsers.pyx:905, in pandas._libs.parsers.TextReader._read_rows()

File parsers.pyx:874, in pandas._libs.parsers.TextReader._tokenize_rows()

File parsers.pyx:891, in pandas._libs.parsers.TextReader._check_tokenize_status()

File parsers.pyx:2061, in pandas._libs.parsers.raise_parser_error()

ParserError: Error tokenizing data. C error: Expected 2 fields in line 11, saw 3
```

```
In [7]: import pandas as pd

ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'

df = pd.read_csv(ruta, sep=',', encoding='utf-8', on_bad_lines='skip')
df.head()

C:\Users\steph\AppData\Local\Temp\ipykernel_13704\1908381003.py:5: DtypeWarning: Columns (1) have mixed types. Specify dtype option on import or set low_memory=False.
      df = pd.read_csv(ruta, sep=',', encoding='utf-8', on_bad_lines='skip')
```

Out[7]:

FECHA PUBLICACIÓN;PAIS PRISIÓN;CONSULADO;DELITO;EXTRADITADO Y O REPATRIADO;SITUACIÓN JURÍDICA;GÉNERO;GRUPO EDAD;UBICACIÓN PAÍS;CANTIDAD;LATITUD;LONGITUD		
17/09/2018;ECUADOR;C. QUITO;NARCOTRÁFICO;DESCONOCIDO;EN JUICIO;MASCULINO;DESCONOCIDO;(-1.831239	-78.183406);45;-1831239;-78183406	
17/09/2018;ITALIA;C. ROMA;OTROS;DESCONOCIDO;EN INVESTIGACIÓN;MASCULINO;DESCONOCIDO;(41.87194	12.56738);1;4187194;1256738	
17/09/2018;ESTADOS UNIDOS;C. NUEVA YORK;EXTORSIÓN;DESCONOCIDO;CONDENADO;MASCULINO;ADULTO;(37.09024	-95.712891);1;3709024;-95712891	
17/09/2018;MEXICO;C. MEXICO;OTROS;DESCONOCIDO;EN INVESTIGACIÓN;MASCULINO;ADULTO JOVEN;(23.634501	-102.552784);7;23634501;-102552784	
17/09/2018;JAPON;C. TOKIO;ROBO / HURTO;DESCONOCIDO;CONDENADO;FEMENINO;ADULTO MAYOR;(36.204824	138.252924);1;36204824;138252924	

```
In [9]: df = pd.read_csv(
    ruta,
    sep=',',
    engine='python',
    encoding='utf-8',
    on_bad_lines='skip',
    quotechar='\"'
)
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 351501 entries, 17/09/2018;ECUADOR;C. QUITO;NARCOTRÁFICO;DESCONOCIDO;EN JUICIO;MASCULINO;DESCONOCIDO;(-1.831239 to 1/03/2025;DESCONOCIDO;C. NEWARK;
LESIONES PERSONALES;DESCONOCIDO;CONDENADO;MASCULINO;DESCONOCIDO;;1;3709024; -95712891
Data columns (total 1 columns):
#    Column
Non-Null Count  Dtype
---  ---
0    FECHA PUBLICACIÓN;PAIS PRISIÓN;CONSULADO;DELITO;EXTRADITADO Y O REPATRIADO;SITUACIÓN JURÍDICA;GÉNERO;GRUPO EDAD;UBICACIÓN PAÍS;CANTIDAD;LATITUD;LONGI
TUD 121772 non-null  object
dtypes: object(1)
memory usage: 5.4+ MB
```

```
In [48]: #corrigiendo dificultades
import pandas as pd

ruta = r'C:\Users\steph\colombianos detenidos exterior.csv'

df = pd.read_csv(
    ruta,
    sep=';', #
    engine='python',
    encoding='utf-8',
    on_bad_lines='skip' # Si hay líneas mal formadas
)

df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 351658 entries, 0 to 351657
Data columns (total 12 columns):
#    Column                                Non-Null Count  Dtype
---  ---
0    FECHA PUBLICACIÓN                    351658 non-null  object
1    PAIS PRISIÓN                        351658 non-null  object
2    CONSULADO                          351658 non-null  object
3    DELITO                             351658 non-null  object
4    EXTRADITADO Y O REPATRIADO          351658 non-null  object
5    SITUACIÓN JURÍDICA                  351658 non-null  object
6    GÉNERO                             351658 non-null  object
7    GRUPO EDAD                         351658 non-null  object
8    UBICACIÓN PAÍS                     121827 non-null  object
9    CANTIDAD                           351658 non-null  int64
10   LATITUD                           309065 non-null  float64
11   LONGITUD                           309065 non-null  float64
dtypes: float64(2), int64(1), object(9)
memory usage: 32.2+ MB
```

```
In [50]: #calcular la media, mediana y desviacion estandar.
import numpy as np

# Extraer la columna CANTIDAD
detenidos = df['CANTIDAD'].dropna().to_numpy()

# Calcular la media
media = np.mean(detenidos)

# Calcular la mediana
mediana = np.median(detenidos)

# Calcular la desviación estándar
desviacion_std = np.std(detenidos)

print("Media de la cantidad:", media)
print("Mediana de la cantidad:", mediana)
print("Desviación Estándar de la cantidad:", desviacion_std)
```

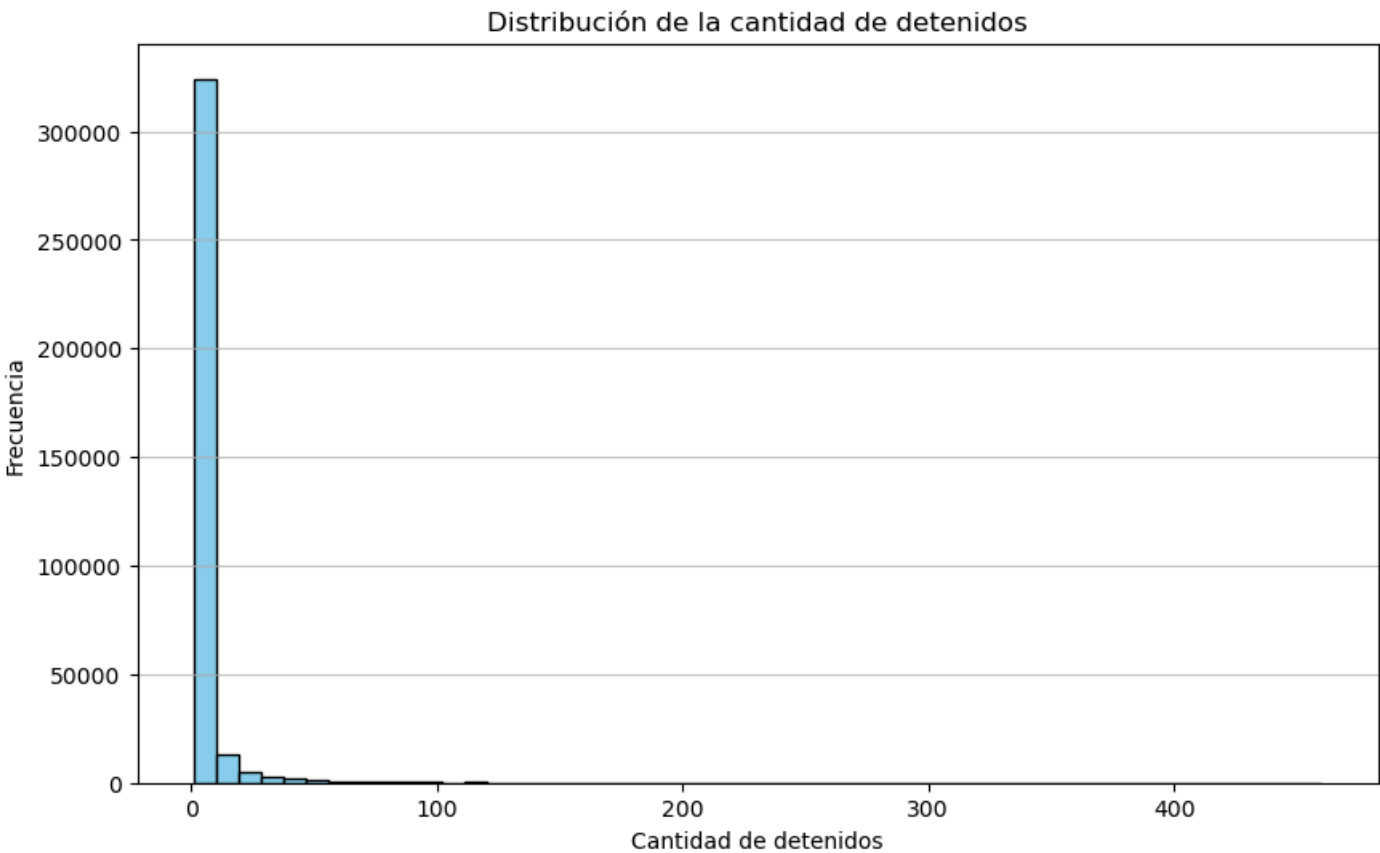
Media de la cantidad: 4.824693309977308
Mediana de la cantidad: 1.0
Desviación Estándar de la cantidad: 17.0100305771676

3. Interpretación de las Medidas Estadísticas

- Media (4.82):
Este valor indica que, en promedio, cada registro reporta aproximadamente 4.82 detenidos. Sin embargo, la media es sensible a valores extremos, por lo que puede estar elevada por algunos registros con cantidades muy altas.
- Mediana (1.0):
La mediana es el valor central de la distribución. En este caso, la mitad de los registros reporta 1 o menos detenidos. Esto sugiere que la mayoría de los registros tienen valores bajos, lo que indica una distribución asimétrica.
- Desviación Estándar (17.01):
Este valor mide la dispersión de los datos respecto a la media. Una desviación estándar tan alta en comparación con la media indica que existe una gran variabilidad en los datos, con registros que tienen cantidades significativamente mayores que el promedio. Esto refuerza la idea de que, aunque la mayoría de los registros son bajos (como indica la mediana), hay algunos registros con valores muy altos que influyen en la media.

```
In [52]: #creando histograma.
import matplotlib.pyplot as plt

# Crear un histograma para la columna "CANTIDAD"
plt.figure(figsize=(10, 6))
plt.hist(df['CANTIDAD'], bins=50, edgecolor='black', color='skyblue')
plt.title("Distribución de la cantidad de detenidos")
plt.xlabel("Cantidad de detenidos")
plt.ylabel("Frecuencia")
plt.grid(axis='y', alpha=0.75)
plt.show()
```



4. Visualización: Histograma de la Cantidad de Detenidos

El histograma anterior muestra cómo se distribuye la columna “CANTIDAD” en la base de datos de colombianos detenidos en el exterior. A partir de la gráfica, se pueden extraer las siguientes observaciones:

- Concentración de Valores Bajos:
La mayoría de los registros se concentran en valores cercanos a 1, lo que concuerda con la mediana de 1 reportada en el análisis estadístico. Esto indica que, en la gran mayoría de los casos, se registra un número reducido de detenidos por fila o evento.
- Presencia de valores atípicos u Outliers:
Aunque los valores bajos dominan la distribución, se observan algunos registros con cantidades notablemente altas (por encima de 100 e incluso superando 400). Estos valores extremos explican la desviación estándar elevada, ya que incrementan significativamente la dispersión de los datos.
- Distribución Marcadamente Asimétrica: La distribución está fuertemente sesgada hacia la derecha. Esto significa que la mayoría de los datos se agrupan en la parte baja, mientras que una minoría de registros se extiende en la cola derecha con valores muy elevados.

4.1. Relevancia para el Problema Inicial:

- Focalización de Casos Atípicos: Identificar y analizar más a fondo los casos con cantidades de detenidos inusualmente altas podría ayudar a entender si se trata de operativos especiales, errores de registro o situaciones puntuales de gran impacto.
- Validación de la Tendencia General: El hecho de que la mayoría de los registros se ubiquen en valores cercanos a 1 sugiere que las detenciones, por lo general, ocurren en eventos de pocos involucrados. Esto puede influir en las estrategias de asistencia consular y en la priorización de recursos.
- El histograma confirma la alta concentración de registros en valores bajos y la existencia de outliers que incrementan la variabilidad de los datos. Esta visualización resulta clave para guiar acciones concretas, ya sea para profundizar en los casos extremos o para diseñar políticas basadas en la tendencia general observada.

5. Patrones, Tendencias y Valores Atípicos

Con base en el análisis estadístico (media, mediana, desviación estándar) y la visualización gráfica (histograma), se pueden destacar los siguientes puntos:

5.1. Concentración de Registros Bajos:

- La mayoría de los valores en la columna "CANTIDAD" se ubican cerca de 1 detenido. Esto se refleja tanto en la mediana (1.0) como en el pico principal del histograma, lo cual sugiere que, en la mayoría de los casos, se registran eventos con un número reducido de detenidos.

5.2. Distribución Asimétrica:

- La diferencia entre la media (4.82) y la mediana (1.0) indica que la distribución está sesgada hacia la derecha. Esto significa que aunque la mayoría de los datos son bajos, existen algunos casos con valores altos que elevan el promedio.

5.3. Valores Atípicos u Outliers:

- El histograma muestra que hay registros con cantidades muy por encima de la mayoría (por ejemplo, por encima de 100 e incluso 400). Estos outliers contribuyen a la elevada desviación estándar (17.01) y pueden representar eventos inusuales o masivos.
- En el contexto de colombianos detenidos en el exterior, estos casos atípicos podrían ser grandes operativos, o podrían indicar inconsistencias en el registro.

5.4. Característica Inesperada:

- La presencia de un número tan elevado de filas con valores de 1 detenido, frente a un pequeño número de filas con cantidades muy altas, confirma que los datos no siguen una distribución “normal” o simétrica. Esto requiere que se tomen precauciones al utilizar medidas sensibles a valores extremos.

In []:

```
In [22]: # Convertir la columna "FECHA PUBLICACIÓN" a datetime
df['FECHA_PUBLICACION'] = pd.to_datetime(df['FECHA PUBLICACIÓN'], dayfirst=True)

# Crear una nueva columna con el valor ordinal de la fecha
df['FECHA_NUM'] = df['FECHA_PUBLICACION'].apply(lambda x: x.toordinal())

# Verificar las primeras filas para confirmar
df[['FECHA PUBLICACIÓN', 'FECHA_PUBLICACION', 'FECHA_NUM']].head()
```

Out[22]:

	FECHA PUBLICACIÓN	FECHA_PUBLICACION	FECHA_NUM
0	17/09/2018	2018-09-17	736954
1	17/09/2018	2018-09-17	736954
2	17/09/2018	2018-09-17	736954
3	17/09/2018	2018-09-17	736954
4	17/09/2018	2018-09-17	736954

In [24]:

```
# Seleccionar las variables y convertirlas.
# Seleccionar FECHA_NUM y CANTIDAD, eliminando valores nulos
x = df['FECHA_NUM'].dropna().to_numpy()
y = df['CANTIDAD'].dropna().to_numpy()

# Verificar La Longitud de ambos arrays
print("Registros en FECHA_NUM:", len(x))
print("Registros en CANTIDAD:", len(y))
```

Registros en FECHA_NUM: 351658
Registros en CANTIDAD: 351658

In [26]:

```
#implementando La regresion lineal.
import numpy as np

# Ajuste Lineal con polyfit
pendiente, intercepto = np.polyfit(x, y, 1)

print("Pendiente (slope):", pendiente)
print("Intercepto:", intercepto)
```

Pendiente (slope): -0.00016941706239899571
Intercepto: 129.90966887714302

6. Interpretación de la Regresión Lineal entre FECHA_NUM y CANTIDAD

Después de convertir la columna "FECHA PUBLICACIÓN" a un valor ordinal (FECHA_NUM) y de seleccionar la columna "CANTIDAD", se aplicó la función `numpy.polyfit(x, y,1)` para obtener la pendiente (slope) e intercepto de la relación lineal, el resultado fue el siguiente:

- Pendiente (slope):(-0.0001694)
 - Este valor negativo, aunque muy pequeño, indica que la cantidad de detenidos tiende a disminuir ligeramente a lo largo del tiempo.
 - En términos prácticos, por cada día que pasa, la cantidad de detenidos disminuiría en aproximadamente 0.000169, lo cual es una variación muy pequeña.
- Intercepto: (129.9097)
 - Este valor representa la cantidad de detenidos que el modelo teóricamente proyecta cuando la fecha ordinal es 0.
 - En la terminos comunes, no suele tener un significado real en este contexto, más allá de ser parte de la ecuación de la recta ajustada.

La pendiente negativa indica una leve tendencia decreciente en la cantidad de detenidos conforme avanza el tiempo, de acuerdo con los datos disponibles.

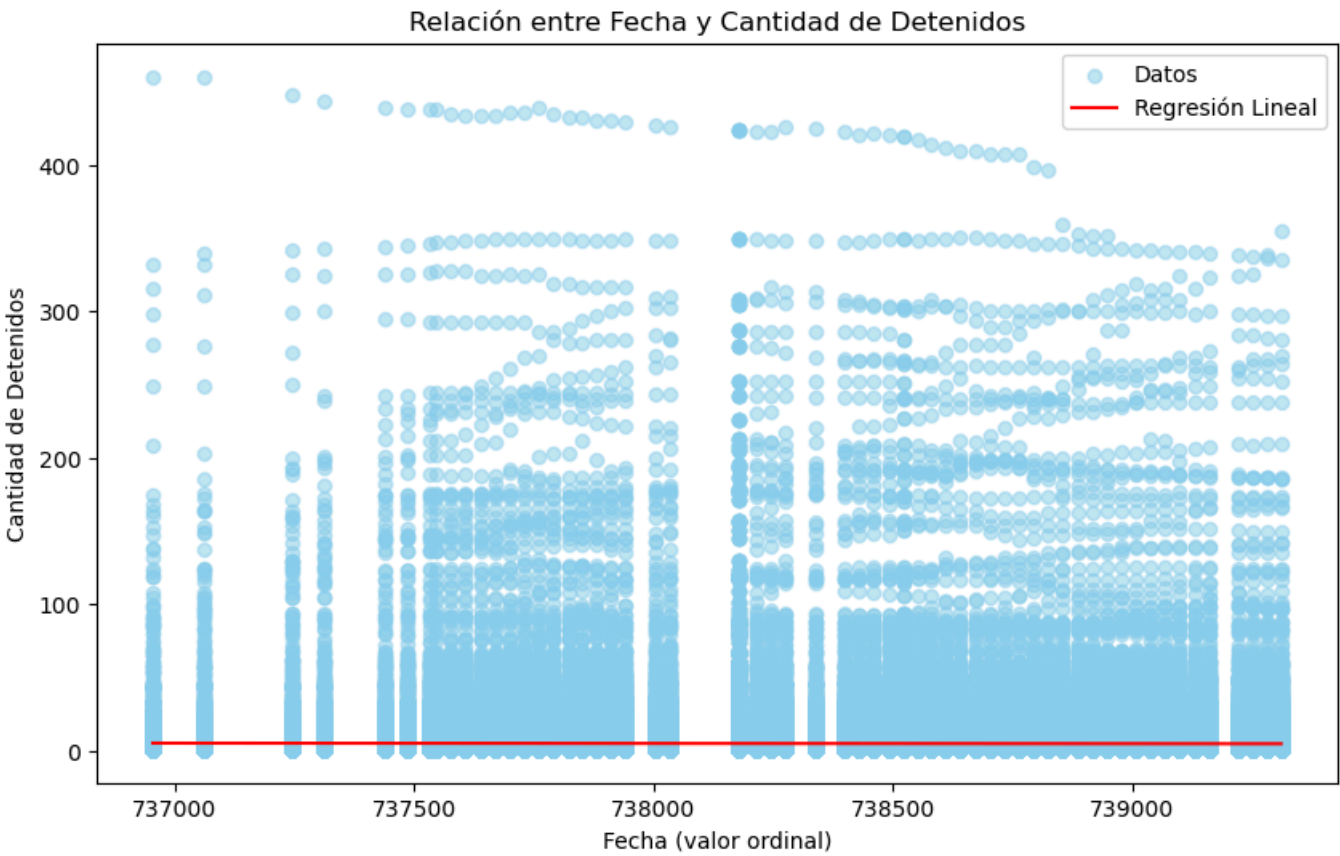
In [36]:

```
#creando La grafica.
import matplotlib.pyplot as plt
import numpy as np

plt.figure(figsize=(10, 6))
plt.scatter(x, y, alpha=0.5, label='Datos', color='skyblue')

# Generar valores para trazar La línea de regresión usando 'pendiente' e 'intercepto'
x_line = np.linspace(min(x), max(x), 100)
y_line = pendiente * x_line + intercepto

plt.plot(x_line, y_line, color='red', label='Regresión Lineal')
plt.title('Relación entre Fecha y Cantidad de Detenidos')
plt.xlabel('Fecha (valor ordinal)')
plt.ylabel('Cantidad de Detenidos')
plt.legend()
plt.show()
```



7. Gráfico de Dispersión con Línea de Regresión

En este gráfico, cada punto azul representa un registro que muestra la relación entre la fecha (eje X, convertida a números) y la cantidad de detenidos (eje Y). La línea roja es la recta que mejor se ajusta a estos puntos, calculada usando una regresión lineal.

8. Conclusiones y tendencias.

- La mayoría de los puntos se encuentran por debajo de 100 detenidos, aunque hay varios casos en los que el número de detenidos llega a superar los 400.
- Esto significa que la cantidad de detenidos varía mucho: la mayoría de los registros son bajos, pero algunos son bastante altos.
- La línea roja se ve casi horizontal, lo que coincide con la pendiente muy pequeña que obtuvimos antes (cerca de cero).
- Esto sugiere que, con el paso de los días (la fecha), la cantidad de detenidos no aumenta ni disminuye de forma marcada; si lo hace, es muy leve.
- Algunos puntos están muy lejos de la línea roja, lo que indica que la fecha por sí sola no explica por qué en ciertos registros hay más o menos detenidos.
- Estos casos podrían ser situaciones especiales o periodos en los que hubo un número anormalmente alto o bajo de detenidos.
- El gráfico confirma que no hay una gran tendencia en la cantidad de detenidos a lo largo del tiempo, o al menos no se ve claramente en un modelo lineal.
- Si queremos entender mejor por qué en ciertos momentos hay más detenidos, es probable que necesitemos incluir otras variables, como el tipo de delito o el país donde sucedió la detención.

9. Posibles Mejoras

Cuando miramos el gráfico y la línea de regresión, se nota que la línea apenas se mueve y no sigue muy bien el comportamiento de los puntos. Esto quiere decir que usar solo la fecha para predecir la cantidad de detenidos no es suficiente para explicar por qué algunos registros tienen más o menos detenidos.

9.1. ¿es suficiente el modelo actual?

- La línea casi plana nos dice que la fecha, por sí sola, no logra explicar los cambios en el número de detenidos.
- Hay muchos puntos que se alejan de la línea, lo que significa que hay otros factores que influyen en esos cambios.

9.2. Mejoras al modelo.

- Agregar más variables Por ejemplo, incluir el tipo de delito, el país o la región donde se realizó la detención, o incluso características demográficas. Así, el modelo tendría más información para explicar la variación en el número de detenidos.
- Implementar un modelo más complejo
En lugar de una simple línea recta, se pueden probar modelos que tengan en cuenta relaciones más complicadas entre las variables, como la regresión múltiple.
- Limpiar y revisar los datos. Es posible que existan registros con errores o valores muy extremos que afecten el modelo. Revisar y limpiar estos datos puede ayudar a obtener un ajuste mejor.

La regresión lineal usando solo la fecha no describe muy bien la relación entre el tiempo y la cantidad de detenidos. Para mejorar la predicción, habría que agregar más información y, posiblemente, usar métodos que capturen relaciones más complejas.

In []:

In []: