

Documentación denuncias:

Este análisis surge de una base de datos del Ministerio de Justicia y Derechos Humanos de la República Argentina. Los datos son utilizados para dar evidencia de la problemática de la perspectiva de género en el país.

Acerca de los datos:

Los datos analizados provienen de una tabla cuya información está recolectada en base a diferentes denuncias de diversas fuentes provistas por el mismo gobierno (denuncias telefónicas, email institucional, entre otras). El período en el que estos datos fueron recolectados comienza en el mes de Enero del año 2021 hasta el mes de Octubre del año 2023.

Algunas de las situaciones registradas son situaciones de pornografía infantil, grooming y explotación sexual comercial de niñas, niños y adolescentes.

Composición de la tabla

La tabla consta de 29 columnas:

"fecha_ingreso", "hora_ingreso", "nro_registro_interno", "situacion", "origen", "es_anonima", "tema", "subtema", "provincia", "localidad", "dependencia_alta", "via_ingreso", "derivacion_institucion", "derivacion_fecha", "derivacion_judicializa", "derivacion2_institucion", "derivacion2_fecha", "derivacion2_judicializa", "derivacion3_institucion", "derivacion3_fecha", "derivacion3_judicializa", "denunciante_nacionalidad", "denunciante_provincia", "denunciante_localidad", "denunciante_tipo", "denunciante_como_conocio_la_linea", "denunciante_genero", "denunciante_edad_aparente", "provincia_indec_id".

La otra tabla que se encuentra dentro del proyecto posee información sobre cada provincia de la República Argentina y su respectiva población. La misma se agregó desde fuentes oficiales facilitados por el gobierno. Dicha tabla fue agregada con la finalidad de poder gestionar las conexiones entre las mismas y darle funciones geográficas a la herramienta de visualización.

Las preguntas que intentamos responder en el informe son enumeradas a continuación:

1. ¿Cuál es el principal motivo de las llamadas?
2. ¿Cómo ha variado la cantidad de llamadas desde la creación de la línea?
3. ¿Cómo se podrá mejorar los medios de contacto para aumentar la eficiencia y resolución de las denuncias?
4. ¿Qué otros tipos de medios se pueden implementar para ayudar a denunciar?
5. ¿Qué provincias son las más afectadas?

Los datos son obtenidos siguiendo los siguientes enlaces y características

- [Datos del Ministerio de Justicia y Derechos Humanos de la República Argentina](#): (Dataset: Dominio público, conjunto de datos disponibles a

través de **Portal de Datos Abiertos de la Justicia Argentina**): Los datos analizados provienen de una tabla cuya información está recolectada en base a diferentes denuncias de diversas fuentes provistas por el mismo gobierno (denuncias telefónicas, email institucional, entre otras). El período en el que estos datos fueron recolectados comienza en el mes de Enero del año 2021 hasta el mes de Octubre del año 2023.

- Datos del sitio oficial del gobierno de la República Argentina. (Datase: Dominio público, conjunto de datos disponibles a través de argentina.gob.ar y en el siguiente [enlace](#)): Los datos informan la demografía de cada provincia según el último censo poblacional.

Los datos utilizados son almacenados de manera local. Pero al ser de fuente abierta se encuentran en los enlaces mencionados anteriormente ya que son pertenecientes al Gobierno de la República Argentina.

Dicha tabla se encuentra organizada de manera vertical, ya que es mejor para la organización. De esta manera pueden observarse de manera más sencilla todos los datos incluso a la hora de realizar las manipulaciones necesarias de la misma tabla.

Respecto al sesgo de los datos no hay problemas y se consideran de gran credibilidad debido a su fuente y a su contenido social mayormente. Son datos de una fuente externa ya que no fueron recolectados por quien los analiza.

Las herramientas escogidas para la limpieza y manipulación de datos de este Dataset serán SQL a través de BigQuery y se utilizó en paralelo R.

La elección está basada en que es una tabla con una cantidad de registros bastante elevada y es necesario procesar estos datos. Las hojas de cálculo en este caso han quedado excluidas por no ser adecuadas para manipular grandes volúmenes de datos.

En cuanto a la limpieza de los datos se tomaron las columnas necesarias para responder las preguntas en las cuales se basa este proyecto. De 29 columnas que dispone el set de datos tomaremos 9, las cuales tienen los registros necesarios. Para llegar a este número de columnas se utilizó BigQuery en el cual se realizó una consulta para extraer una tabla con las columnas necesarias y se descargó en el almacenamiento local en la carpeta del proyecto. También se procede a dar un orden por fechas desde el año 2020, que fue el año de inicio de la toma de datos, hasta su fin en el año 2023.

Respecto de algunas columnas se han modificado los valores nulos a un valor de tipo "string" ya que son importantes para las métricas y la información que se desea compartir. Mediante análisis no es necesario cambiar el tipo de ninguno de los datos ya que cada columna corresponde a su tipo de dato.

Partiendo así del archivo que se utilizó para el análisis, el cual se encontraba en formato csv y transformado a un archivo .xlsx y se puede acceder mediante el siguiente [enlace](#).

Para la limpieza con R se utilizaron los paquetes **tidyverse**, **dployr**, **ggplot**, entre otros los cuales fueron instalados y luego cargadas las respectivas librerías.

Luego para realizar el informe y visualizaciones se procede a operar con Power BI. En dicha herramienta se genera un informe respondiendo a las preguntas propuestas y realizando las conclusiones correspondientes. De manera extra se genera un Dashboard para dar una vista general del análisis realizado.