UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escuela Técnica Superior de Ingeniería Informática

Universidad Politécnica de Valencia

Tormenta: una plataforma de centralización y consulta de datos abiertos sobre violencia de género en España

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autora: Andrea Natalia Naranjo Chávez

Tutora: Ángeles Calduch Losa

2020-2021

Tormenta: una plataforma de centralización y consulta de datos abiertos sobre violencia de género en España

Dedicatoria

Para mi madre, Beatriz, que escuchaba las canciones de Tormenta cuando yo era una niña. Ahora tiene otras favoritas.

Agradecimientos

A mi familia tan amada que me espera hasta que cumpla mis sueños. ¿Cuántos sueños? me preguntan desde el otro lado del mar. A David, porque Tormenta no existiría sin su apoyo sin condición y su compañía. A Marta por las lecturas y los consejos. A Ángeles por hacer que este proyecto tenga múltiples caminos. A Critical Switch por las complicidades, el cariño y la explosión de búsquedas comunes. A Punset y a Sagan por los ronroneos.

Resumen

Vivir sin violencia de género es un derecho fundamental. Sin embargo, de acuerdo con la ONU, 1 de cada 3 mujeres la ha vivido en su expresión física o sexual, por eso 2 de los 17 Objetivos de Desarrollo Sostenible (ODS) incluyen su erradicación como una meta a conseguir antes del año 2030. Como apoyo a este objetivo y haciendo uso de los datos abiertos, entendidos como un recurso público, se ha desarrollado *Tormenta* como una plataforma de centralización, normalización y consulta, con la intención de que sus usuarios puedan realizar análisis en conjunto y compartir datos que actualmente se encuentran dispersos o sin relaciones entre ellos, a pesar de las múltiples fuentes de datos que hay en España sobre este tema. Tormenta contiene una selección de datos abiertos del Consejo General del Poder Judicial (CGPJ) español entre los años 2015 y 2019 y datos poblacionales del Instituto Nacional de Estadística (INE). Haciendo uso de la tecnología de gestión, los datos han sido tratados en 3 etapas: extracción, transformación y carga de datos a través de secuencias de comandos creadas con JavaScript.

Palabras clave: Centralización de datos, Datos abiertos, Proceso de ETL, ODS, Violencia de género, Análisis de datos.

Abstract

To live without gender violence is a fundamental right. However, in accordance with the UN, 1 in every 3 women has experienced it in its physical or sexual expression, for that reason, 2 out of 17 Sustainable Development Goals (SDGs) include its eradication as a goal to achieve before year 2030. As a support to this goal and making use of the open data, understood as a public resource, *Tormenta* has been developed as a centralization, normalization and consultation platform with the purpose of allowing its users to perform joint analysis and share data that is currently scattered or without relations between them, despite the multiple data sources that Spain has. Tormenta contains a set of open data from the Spanish Consejo General del Poder Judicial (CGPJ) between the years 2015 and 2019 and population data from the Instituto Nacional de Estadística (INE). Making use of the management technology, the data has been treated in 3 steps: extraction, transformation and data load through command sequences created with JavaScript.

Keywords: Data centralization, Open data, ETL process, ODS, Gender violence, Data analysis.

Tabla de contenidos

Lista de figuras

Lista de tablas

Gráficas

Acrónimos

BD: Base de datos.

CGPJ: Consejo General del Poder Judicial.

ETL: por sus siglas en inglés Extract, Transform and Load. En español: extraer, transformar y cargar.

ONU: Organización de Naciones Unidas.

ODS: Objetivos de Desarrollo Sostenible.

TFM: Trabajo Final de Máster.

SGBD: Sistemas de gestión de bases de datos.

WEB: Abreviatura para World Wide Web.

1.Introducción

En un mundo en el que se dice que
la igualdad es real, para desenmascararla
nada mejor que las cifras... (Varela, 2004, 168)

La violencia de género es un problema que ocurre dentro de los espacios íntimos, pero es un tema público porque afecta a la mitad de la población. Los esfuerzos por eliminarla son necesarios para alcanzar una sociedad más igualitaria y los datos abiertos son la herramienta que se utiliza en este proyecto para apoyar este objetivo mundial. En España existen múltiples fuentes de datos abiertos sobre la violencia de género, algunos se utilizan en proyectos que buscan evidenciar el problema, pero la falta de relaciones entre estos datos limita las posibilidades de profundizar en los análisis. Por otra parte, se utilizan los procesos de gestión de datos, ya que han tenido un rápido avance desde su aparición y ayudan a crear las relaciones que se necesitan para realizar estudios más profundos con los datos. A continuación se describen los datos abiertos, la gestión de datos, las fuentes de datos y, finalmente, la estructura del TFM.

Los datos del CGPJ son datos de procesos legales por lo que no dan una dimensión real del alcance de la violencia. Estos datos se recogen cuando se ha iniciado una acción legal por parte de la víctima o de un familiar, además de lo que puede conllevar esta decisión en la víctima y la familia, está la limitación con la que cuentan mujeres extranjeras sin documentos legales para residir en el país o que dependen de sus convivientes para hacerlo.

Otro problema es que estos datos, que nos pueden dar una dimensión de la violencia, se encuentran dispersos o sin establecer relaciones entre ellos. Por ejemplo, en el Portal Estadístico de la Delegación del Gobierno contra la Violencia de Género se encuentran datos sobre denuncias, llamadas por violencia de género al 016 o usuarias de ATENPRO, entre otros, pero las consultas se hacen por fuente de datos, siendo difícil usar este portal como herramienta para responder a preguntas tales como: “¿Tenemos aproximadamente los mismos casos de violencia de género en función de la población en cada provincia?” “¿Hay relación entre la renta media y la violencia de género?”.

Sin embargo, el problema de la dispersión de datos se podría resolver con la gestión integrada de estos. De acuerdo con Gerardo (2008, 5) desde la aparición de la gestión de datos por computador este proceso se ha visto acelerado brindando nuevas soluciones, pues son sistemas cada vez más robustos y fiables, y disponen de lenguajes e interfaces de acceso y manipulación cada vez más amigables. Para los sistemas de bases de datos relacionales o SGBD se utilizan procesos ETL, estos son desarrollados e investigados sobre todo desde el ámbito empresarial, por lo que se encuentran en las soluciones de inteligencia de negocios.

En el presente trabajo se hace uso de estos avances para ayudar a conseguir el objetivo que persigue Tormenta, demostrar las posibilidades que ofrece la centralización de datos abiertos y el establecimiento de relaciones entre las diferentes fuentes de datos fuera del ámbito de los negocios. Para hacerlo se utiliza Redash, una herramienta *Open Source* que permite consultar los datos y generar visualizaciones. De esta forma, las secuencias de comandos junto con Redash permiten analizar los datos del CGPJ sin que sus usuarios deban preocuparse por su tratamiento. A través de esto, se busca mostrar las ventajas de la centralización de datos abiertos y la profundidad del análisis que se puede generar al establecer relaciones entre distintas fuentes de datos. Para probar la utilidad de esta herramienta se realiza un análisis estadístico en la Comunidad Valenciana.

Las fuentes de datos con las que se va a trabajar en este TFM son los datos cuantitativos proporcionados por el CGPJ de los últimos 5 años disponibles, que comprenden el periodo desde el año 2015 hasta 2019 y los datos poblacionales del INE. Se ha escogido el CGPJ tanto por la extensa información que poseen sus datos como porque se han encontrado al menos 2 proyectos que los utilizan, además de porque estos datos permiten trabajar con una separación por provincias y por comunidades autónomas, por lo que se puede diferenciar los niveles de violencia en relación a la cantidad de habitantes y no solamente al número de denuncias.

Este trabajo se desarrolla en 6 capítulos. Comienza con la motivación personal y la necesidad de trabajar con datos de la violencia de género. Continúa el contexto tecnológico en el que se describen otros proyectos web que han utilizado datos abiertos y, que a la vez, han servido de análisis para la propuesta de Tormenta.

Después, la descripción de la solución propuesta, el diseño y la tecnología utilizada. Más tarde, para demostrar el potencial de Tormenta, se realiza un análisis estadístico. Finalmente, se cierra con las conclusiones, el desarrollo de este trabajo en relación a los conocimientos personales y técnicos necesitados vinculados con las materias cursadas en el MUGI y la exposición de trabajos futuros que se podrían realizar con Tormenta.

1.1. Motivación

La violencia de género ocurre. La vivió mi tía con su exesposo. La vivió mi amiga española con su exnovio. La vivió mi prima con su marido. La vivía la señora que lavaba mi ropa cuando mi madre estudiaba. La vivió la famosa escritora Kameron Hurley. La vivió Mónica Ojeda, una de las escritoras latinoamericanas más influyentes de 2018. La vivió Angie Carrillo[1] antes de ser asesinada por su exnovio. Para descubrir las vivencias de mujeres cercanas a mí y generar estas conversaciones han tenido que pasar varios años. Conversaciones que empiezan incómodas y luego como una suerte de confesión, de reconocimiento o no. Incluso después de hablar, algunas mujeres la siguen viviendo, porque se ven juzgadas por una sociedad que cree que la violencia ocurre pocas veces y en circunstancias desfavorecidas económicamente. Otras veces, la violencia verbal, psicológica o patrimonial es entendida como otra representación del amor.

La primera propuesta informal de este TFM fue visibilizar la violencia de género con un análisis estadístico de los datos recolectados por el CGPJ, en los últimos 10 años. Tras encontrar varios análisis con los mismos datos y otros problemas de tratamiento de datos, Tormenta, el nombre del proyecto actual, es el nuevo enfoque del tema. A continuación describo los problemas que se volvieron motivos para crear Tormenta. También escribo de los datos abiertos y de su uso dentro del periodismo de datos.

Al iniciar el anterior proyecto encontré portales web que muestran gráficas con análisis de 10 años de los datos, incluso alguno desde 2004. Los análisis cuentan con gráficas interactivas que permiten establecer comparaciones, por ejemplo, entre provincias o entre años. Sin embargo, hay limitaciones para generar nuevos análisis. No se pueden

[1] El caso de Angie Carrillo es el primer caso que marca el reconocimiento del Femicidio en el Ecuador después de que el asesino confesara su crimen. El feminicidio aún no es reconocido en este país.

crear otras gráficas a partir de las relaciones entre las tablas porque las relaciones no están establecidas. Para obtener análisis propios me acerqué a las fuentes de datos originales, encontré datos en formatos PDF o en XLS. De acuerdo con el esquema[2] de 5 estrellas de los datos abiertos, desarrollado por Tim Berners-Lee, este tipo de formatos está catalogado con 1 y 2 estrellas por los altos costos y bajos beneficios que tienen al ser tratados.

Los datos abiertos son datos que se disponen desde las instituciones públicas, de acuerdo con Ruvalcaba Gómez (2020) son un medio para generar transparencia, impulsar su uso en nuevas iniciativas que mejoren la calidad de vida de sus ciudadanos, o que generen beneficios económicos. En España, de acuerdo con el mismo autor, los datos abiertos se han incrementado desde "la entrada en vigor de la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, que se publicó en el Boletín Oficial del Estado el 10 de diciembre de 2013". De ahí que los datos abiertos son un bien público, por lo que muchas veces están vinculados con la transparencia de los gobiernos, no obstante también con las iniciativas que surgen sin restricciones por el tipo de uso, es decir, que estos pueden ser sociales y económicos, entre otros. En el caso de Tormenta, se entiende como un recurso del que se hace uso con beneficios para la ciudadanía.

¿Quiénes utilizan estos datos? Además de la sociedad a partir de las comunidades, están los periodistas, sobre todo del periodismo de datos. Esta es una vertiente reciente, si bien no en Estados Unidos, sí en España[3] y Latinoamérica, tomando en cuenta que el primer Congreso sobre este tema se realizó en 2016 en Málaga. El periodismo de datos es una especialización que se va extendiendo, porque existe una necesidad, Mar Cabra en el prólogo de (Blanco Castilla & Quesada, 2016, 10) lo dice así "... empezar a usar datos en bruto para encontrar temas que ocurrían de manera sistemática y así descubrir asuntos que pasan desapercibidos en el día a día o que requieren del uso de ordenadores para ser entendidos". De ahí que en este TFM se plantea un producto mínimo viable que planea crecer y, como trabajo a futuro, ser una herramienta para periodistas que visibilicen la violencia de género.

[2] Tim Berners-Lee el esquema de las 5 estrellas para los datos abiertos. Aquí se resumen los beneficios y costes de los datos en cada una de estas estrellas. Disponible en: https://5stardata.info/es/#costs-benefits

[3] De acuerdo con la introducción de Mar Cabra en Blanco Castilla & Quesada, (2016, p.9) el primer evento en el que se habla de "tener un debate público y acelerar el desarrollo del Periodismo de datos en España" sucedió el 15 de febrero de 2011 en Medialab-Prado.

Para concluir, el tiempo para elaborar reportajes en las redacciones está definido por la cantidad de periodistas y la premura de los temas. Los análisis profundos se permiten siempre que tomen el menor tiempo posible. Y para organizaciones más pequeñas, que cuentan con el apoyo de activistas, generar estos análisis es aún menos accesible. Tener la información centralizada posibilitará análisis más profundos con menor inversión de tiempo. Estos análisis no son necesariamente desde una perspectiva de *Big Data*, sino del uso estratégico para resolver problemas más locales como los de una provincia, una comunidad autónoma o un municipio.

1.2. Objetivos

Tormenta es una plataforma de centralización, normalización y consulta de datos abiertos sobre la violencia de género. Su objetivo principal es demostrar las posibilidades que ofrece la centralización de datos abiertos y el establecimiento de relaciones entre las diferentes fuentes de datos. Para hacerlo, los objetivos específicos que debe cumplir son:

- Normalizar y centralizar los datos abiertos.
- Establecer relaciones entre datos de fuentes diferentes.
- Ofrecer un método de acceso y consulta a los datos centralizados.
- Demostrar su utilidad a través de diversos análisis estadísticos.

Los datos abiertos que se utilizarán para demostrar las posibilidades descritas en el objetivo general, son: datos del CGPJ referentes a violencia de género (de ahí que el nombre del actual TFM incluya información sobre el ámbito de los datos). También se utilizarán los datos de las poblaciones por provincias y a su vez agrupadas por comunidades.

1.3. Impacto esperado

Se espera que esta herramienta sea utilizada por la sociedad y por periodistas de datos para disminuir su tiempo de análisis de datos sobre la violencia de género. De esta forma se pretende dar más visibilidad a la problemática. Además, se quiere contribuir con la Agenda Mundial que marcan los ODS y la eliminación de este tipo de violencia. El uso de estos datos tratados y relacionados, además de ser un aporte para

la sociedad por la posibilidad de experimentación, brinda la oportunidad de expansión, uso, estudio y mejora por estar disponibles sin licencias restrictivas.

A través de Tormenta se cumplen 2 de los 17 objetivos de ODS. El primero es el objetivo número 5 sobre la "Igualdad de género" que especifica en el apartado 5.2 la necesidad de "Acabar con la violencia contra las mujeres". El segundo objetivo es el 16 sobre la "Paz, justicia e instituciones sólidas", que en el apartado 16.1 propone la "Reducción de la violencia". Sobre el aporte a la sociedad, como comentó Renata Ávila en su reciente charla dentro del Decidim Fest:

> "Tenemos derecho a que esos datos que están generados por la ciudad y capturados y almacenados por la administración pública, sean públicos... que sean este material que nos va a permitir a todos y todas y no solo a las compañías que absorben esos datos, abrir procesos de creación, de experimentación, de investigación y de exploración del espacio en el que vivimos. Es acceso a un bien público fundamental que nos va a permitir hacer nuestras ciudades mejores." (Ávila, 2020)

En búsqueda de este acceso público, Tormenta estará disponible a través de una dirección web, podrá ser usada libremente, su código estará en un repositorio de GitLab para que pueda ser, como se ha dicho anteriormente, usado, modificado, estudiado y mejorado según las 4 libertades del software libre.

Además de la sociedad, se espera que perfiles como los de periodistas de datos accedan a la herramienta para disminuir el tiempo destinado a los procesos previos al análisis. Por una parte pueden acceder usuarios con conocimientos de SQL, porque podrán generar consultas propias que estarán disponibles para el resto de usuarios. Por otra parte, cualquier usuario puede acceder a las consultas iniciales que se han creado y pueden interactuar con ellas a través de filtros por año, provincia o comunidad, también generar visualizaciones propias haciendo uso de las relaciones entre las diferentes tablas. Hay que especificar que se deben crear los usuarios y definir los accesos. Esto hace que quien administra Redash con la BD de Tormenta tenga control de que no se ha modificado la BD de la que se hace uso.

1.4. Metodología

En este proyecto se parte del proceso ETL para trabajar los datos, siguiendo las 3 etapas: extracción, transformación y carga de datos en una BD relacional. Para

empezar se definen las fuentes de datos, después las etapas se realizan con varias secuencias de comandos. Además del proceso, las fuentes y de qué es una secuencia de comandos, también se describe Docker y Docker-Compose, las herramientas utilizadas para unir las diferentes tecnologías en un único paquete, lo que facilita su réplica sin necesidad de instalar dependencias externas.

Las fuentes de información públicas desde las que se extraen los datos de prueba son 2, en trabajos futuros se espera contar con la colaboración de los usuarios y ampliar estas fuentes de información, entonces, se obtienen los datos del CGPJ y datos poblacionales del INE. En el marco referencial se detalla la historia de la recolección de datos, su importancia y la frecuencia con la que son publicados. Desde estas fuentes se inicia el proceso ETL.

El proceso ETL es muy conocido en el mundo del tratamiento de datos y de acuerdo con Bustamante et al. (2013, p. 185) se resume en 3 pasos:

- El primero “analizar las fuentes de datos existentes para encontrar la semántica oculta en ellas”.
- El segundo “diseñar el flujo de trabajo que extraiga los datos desde las fuentes, repare sus inconsistencias, los transforme en un formato deseado” que constituye el proceso más largo. Será resuelto con la secuencia de comandos desarrollados con JavaScript.
- Finalmente, “los inserte en la bodega de datos.”, en Tormenta será en una BD relacional.

Sobre el segundo punto, la secuencia de comandos es un código propio que da flexibilidad para crear una librería con diferentes extracciones y transformaciones. Estas funcionalidades se pueden ir enriqueciendo y tener un mayor alcance, por ejemplo ampliar la lista de fuentes o formatos soportados. La secuencia de comandos también permite hacer procesos personalizados como la normalización de los nombres de las provincias y comunidades en los casos en lo que se tienen varias formas de nombrarlas. Las secuencias de comandos serán creadas con JavaScript porque es uno de los lenguajes de programación más extendidos, lo que aumenta las posibilidades de reutilización fuera de Tormenta o para incorporar nuevas fuentes en

trabajos futuros. El código estará disponible en un repositorio público de GitLab, un servicio web especializado en el control de versiones y trabajo en equipo.

El tercer punto implica el uso de una BD. Para Celma Giménez et al. (2002, p.49) “Una BD es una colección de datos estructurados de acuerdo a las reglas de un modelo” en el caso de las BD relacionales es el modelo relacional, como aclaran los autores, el modelo fue propuesto en 1970 y tuvo éxito por su sencillez. Este tipo de BD forma un conjunto de información organizada que posee relación entre sí. Contiene una colección de tablas que almacenan información de forma estructurada. Cada entidad posee un atributo o conjunto de atributos que la hace única en la BD. Puesto que las BD deben ser manipuladas a través de Sistemas de Gestión de Bases de Datos (SGBD), para el modelo relacional la consulta se realiza a través del lenguaje SQL, así se obtiene el potencial de las relaciones establecidas.

Por su parte, Docker, de acuerdo con su página web, es una herramienta de código abierto que se lanzó en 2013. Permite empaquetar el código con sus dependencias para que la aplicación o herramienta disponga de lo necesario para ejecutarse. Es decir, aísla el software en su entorno. Lo que lo diferencia de las máquinas virtuales es que virtualiza el sistema operativo, en lugar del hardware, por lo que se vuelve más eficiente, además de que ocupa menos espacio, sin contar con los tiempos de levantar una máquina virtual con el software preinstalado. Docker-Compose es la herramienta que permite manejar múltiples contenedores de Docker.

1.5. Convenciones

- El código fuente y las referencias a sus funciones o clases se muestran en letra Source Sans Pro, con grosor normal. Y sólo se emplea esta tipología para este tipo de contenido.
- Las palabras extranjeras se marcan en cursiva.
- También se marcarán en cursiva los nombres de los ficheros o archivos que intervienen en el desarrollo de la secuencia de comandos de Tormenta.
- Se entrecomillan las citas textuales externas a la obra.

2. Contexto

Para empezar, se pondrá en contexto dos proyectos que han sido creados hace menos de 2 años, se encuentran finalizados y en una fase extra de recolección de datos. Las similitudes entre ellos son: el apoyo del Ayuntamiento de Barcelona, los datos que utilizan son datos públicos sobre la violencia de género, posiblemente comparten fuentes de información, y su objetivo final es promover el análisis de los datos abiertos para visibilizar la problemática.

El primer proyecto es *Datos contra el Ruido*[4] de la asociación Digital Fems[5] que también cuenta con el apoyo de DataPlace, StoryData, School of Feminism, Soko.tech. El proyecto es una plataforma que "muestra la visualización de datos acerca de las diferentes tipologías de violencia: de género, doméstica y sexual." (Digital Fems, n.d.). En su página web, dentro de los 7 apartados que conforman el proyecto, la información se encuentra en el primero, llamado: Datos. En la siguiente figura se muestra la portada del proyecto y sus apartados:

Figura 2.1 Apartados web del proyecto Digital Fems

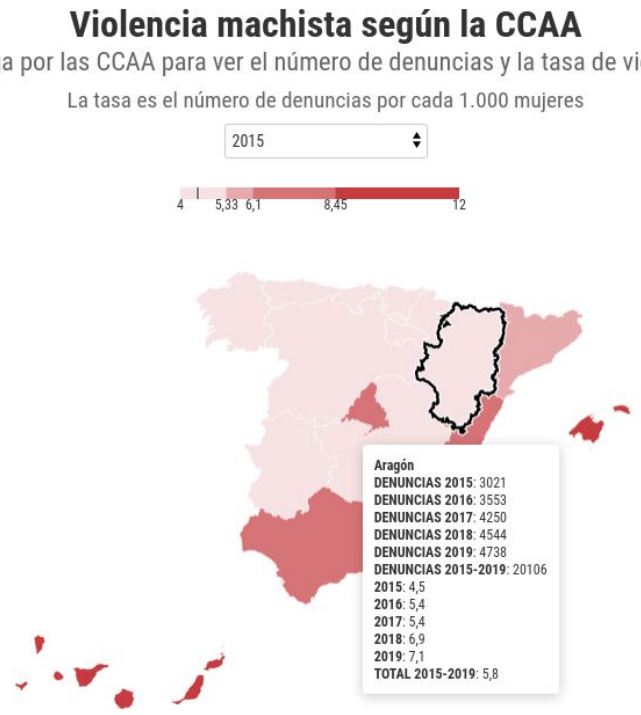
Fuente: Portal web de Datos contra el ruido

[4] "Datos contra el ruido publicó en mayo de 2020 más de 25 visualizaciones y 40 *datasets*." Recuperado de: https://www.goteo.org/project/datoscontraelruido2

[5] Digital Fems se define como una asociación que apoya en proyectos que acaben con los estereotipos de género. Recuperado de: http://www.digitalfems.org/

El apartado Datos cuenta con gráficas generadas por Flourish[6]. Se pueden hacer búsquedas interactivas, sobre todo aplicar filtros por: año, comunidad autónoma, tipos de delitos y tipos de agresores de acuerdo a la violencia ejercida y el total de denuncias que acaban en sentencia condenatoria. Se muestra un ejemplo en la siguiente figura permite elegir años y comunidades autónomas. También acceder al total de 2015 a 2019.

Figura 2.2 Gráfica interactiva con datos de denuncias y la tasa de violencia



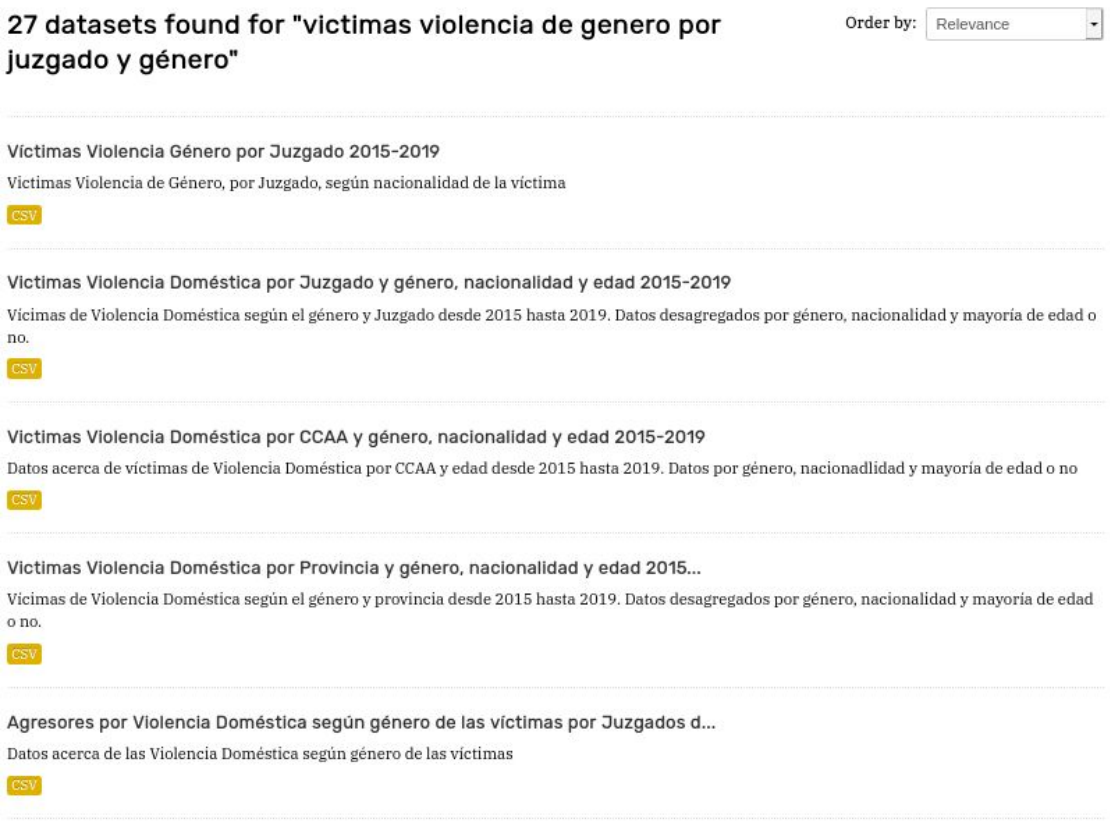
Fuente: Portal web de Datos contra el ruido

Por un lado, las gráficas limitan la posibilidad de realizar otras comparaciones. Es decir, si en un análisis se quiere establecer o encontrar correlaciones entre la cantidad de población de una comunidad autónoma y la cantidad de delitos, las gráficas no brindan esta información. Para hacerlo debería acceder al apartado: Gender Data Lab. Este apartado redirige a otra página web que es un repositorio de conjuntos de datos que han sido tratados y se encuentran en formato CSV[7]:

[6] Este software permite generar gráficas a partir de datos procesados previamente. Página oficial de Fluorish: https://flourish.studio/blog/

[7] Comma-separated values o CSV es un tipo de archivo con datos separados por comas, cada registro se encuentra una línea. (Wikipedia, n.d.)

Figura 2.3 Conjuntos de datos en formato CSV de Víctimas Violencia Doméstica por Juzgado y género, nacionalidad y edad 2015-2019

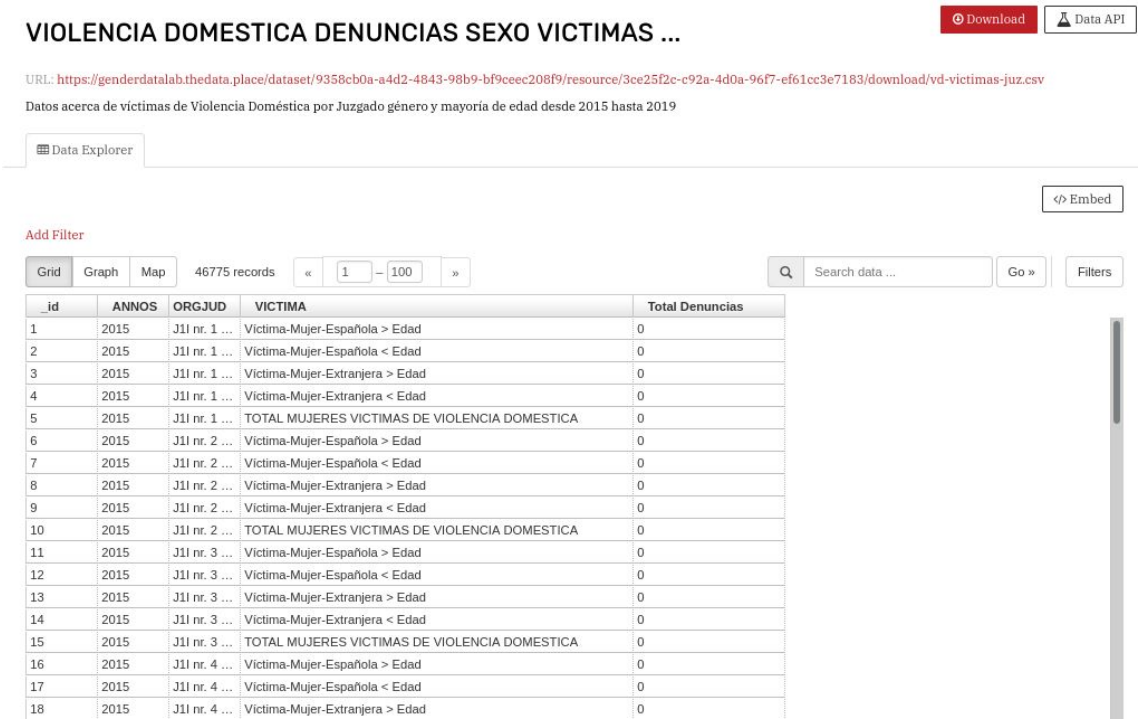


Fuente: Gender Data Lab de Digital Fems https://cutt.ly/IhjcowS

El contenido de los archivos CSV son tablas específicas. La siguiente figura muestra la tabla de datos de "Violencia Domestica Denuncias Sexo Vícitimas..."[8] Los datos se encuentran por juzgado, por año y por tipo de víctima de acuerdo a la nacionalidad. Una persona usuaria de estos datos debería establecer las relaciones entre esta y otras tablas que contienen más información como: comunidades autónomas, provincias, población, denuncias, sentencias.

[8] Estos datos han sido tomados del proyecto Gender Data Lab y se encuentran disponibles en: https://genderdatalab.thedata.place/dataset/victimas-violencia-domestica-por-juzgado-y-genero-nacionalidad-y-edad-2015-2018/resource/3ce25f2c-c92a-4d0a-96f7-ef61cc3e7183. Último acceso 24/11/2020

Figura 2.4 Conjunto de datos con id, años, juzgados, víctima española o migrante y total de denuncias.



VIOLENCIA DOMESTICA DENUNCIAS SEXO VICTIMAS ...

Download | Data API

URL: https://genderdatalab.thedata.place/dataset/9358cb0a-a4d2-4843-98b9-bf9ceec208f9/resource/3ce25f2c-c92a-4d0a-96f7-ef61cc3e7183/download/vd-victimas-juz.csv

Datos acerca de víctimas de Violencia Doméstica por Juzgado género y mayoría de edad desde 2015 hasta 2019

Data Explorer | Embed

Add Filter

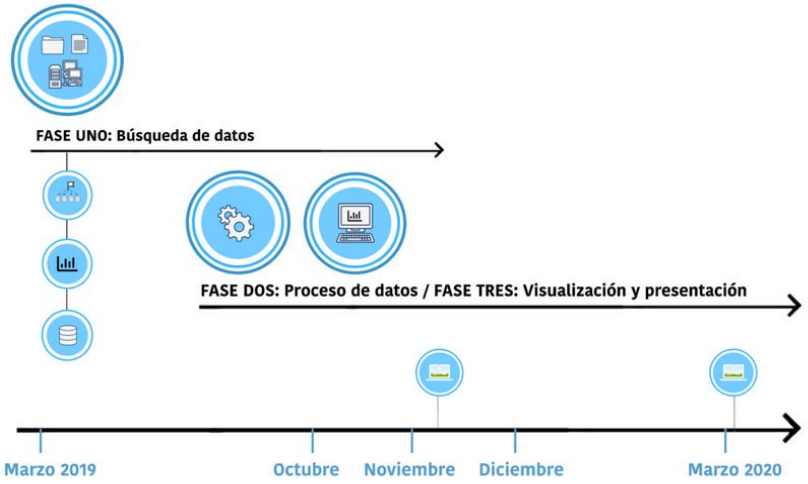
Grid | Graph | Map | 46775 records | « 1 – 100 » | Search data ... | Go » | Filters

_id	ANNOS	ORGJUD	VICTIMA	Total Denuncias
1	2015	J1l nr. 1 ...	Víctima-Mujer-Española > Edad	0
2	2015	J1l nr. 1 ...	Víctima-Mujer-Española < Edad	0
3	2015	J1l nr. 1 ...	Víctima-Mujer-Extranjera > Edad	0
4	2015	J1l nr. 1 ...	Víctima-Mujer-Extranjera < Edad	0
5	2015	J1l nr. 1 ...	TOTAL MUJERES VICTIMAS DE VIOLENCIA DOMESTICA	0
6	2015	J1l nr. 2 ...	Víctima-Mujer-Española > Edad	0
7	2015	J1l nr. 2 ...	Víctima-Mujer-Española < Edad	0
8	2015	J1l nr. 2 ...	Víctima-Mujer-Extranjera > Edad	0
9	2015	J1l nr. 2 ...	Víctima-Mujer-Extranjera < Edad	0
10	2015	J1l nr. 2 ...	TOTAL MUJERES VICTIMAS DE VIOLENCIA DOMESTICA	0
11	2015	J1l nr. 3 ...	Víctima-Mujer-Española > Edad	0
12	2015	J1l nr. 3 ...	Víctima-Mujer-Española < Edad	0
13	2015	J1l nr. 3 ...	Víctima-Mujer-Extranjera > Edad	0
14	2015	J1l nr. 3 ...	Víctima-Mujer-Extranjera < Edad	0
15	2015	J1l nr. 3 ...	TOTAL MUJERES VICTIMAS DE VIOLENCIA DOMESTICA	0
16	2015	J1l nr. 4 ...	Víctima-Mujer-Española > Edad	0
17	2015	J1l nr. 4 ...	Víctima-Mujer-Española < Edad	0
18	2015	J1l nr. 4 ...	Víctima-Mujer-Extranjera > Edad	0

Fuente: Gender Data Lab de Digital Fems https://cutt.ly/8hsai7H

El segundo proyecto se llama Datos x Violencia x Mujeres. Como se puede observar en la figura 2.5, el proyecto inició en marzo del 2019 y finalizó un año después. El objetivo fue ser un "portal interactivo (sic) que permitirá a la ciudadanía comparar y contrastar datos" (Barcelona Iniciativa Open Data, n.d.). En el mismo portal se encuentran las principales fuentes de información utilizadas: "el Observatorio contra la Violencia de Género y Doméstica del Consejo General del Poder Judicial, la Delegación de Violencia Género del gobierno español y los organismos autonómicos dedicados a la Igualdad. En cuanto las cuestiones sobre recursos públicos, se han consultado los presupuestos generales del Estado y los de cada CCAA." Hay que aclarar, que de acuerdo con la información en la página principal, la búsqueda de datos continúa.

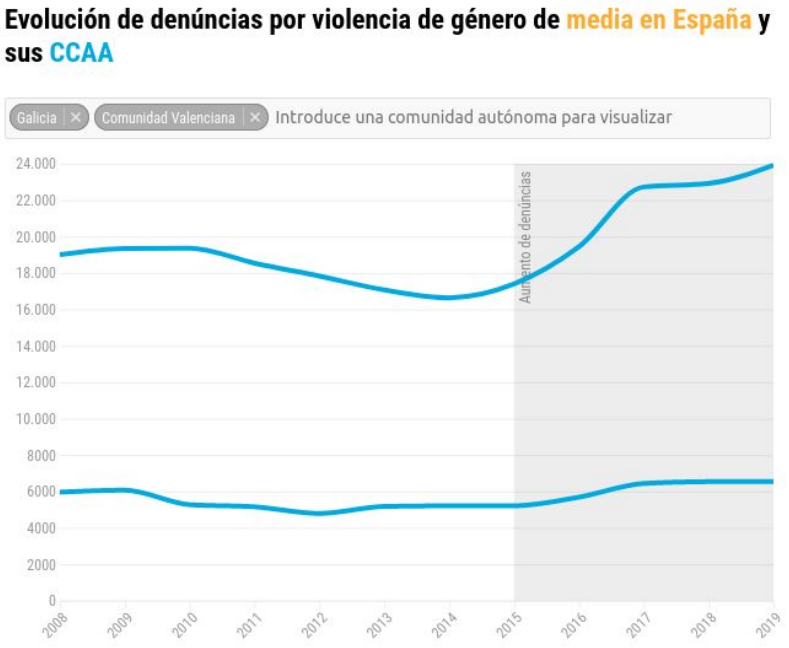
Figura 2.5 Fases del proyecto Datos x Violencia x Mujeres



Fuente: Iniciativa Barcelona Open Data

Además de la gráfica interactiva, que al igual que en *Datos contra el ruido* han sido creados con Flourish, se pueden descargar los datos que la generan. Estos también se encuentran en formato CSV. En la siguiente figura se muestra una captura de pantalla de la gráfica interactiva en la que se pueden contrastar los datos entre distintas comunidades autónomas por año, tomando en cuenta el número de denuncias:

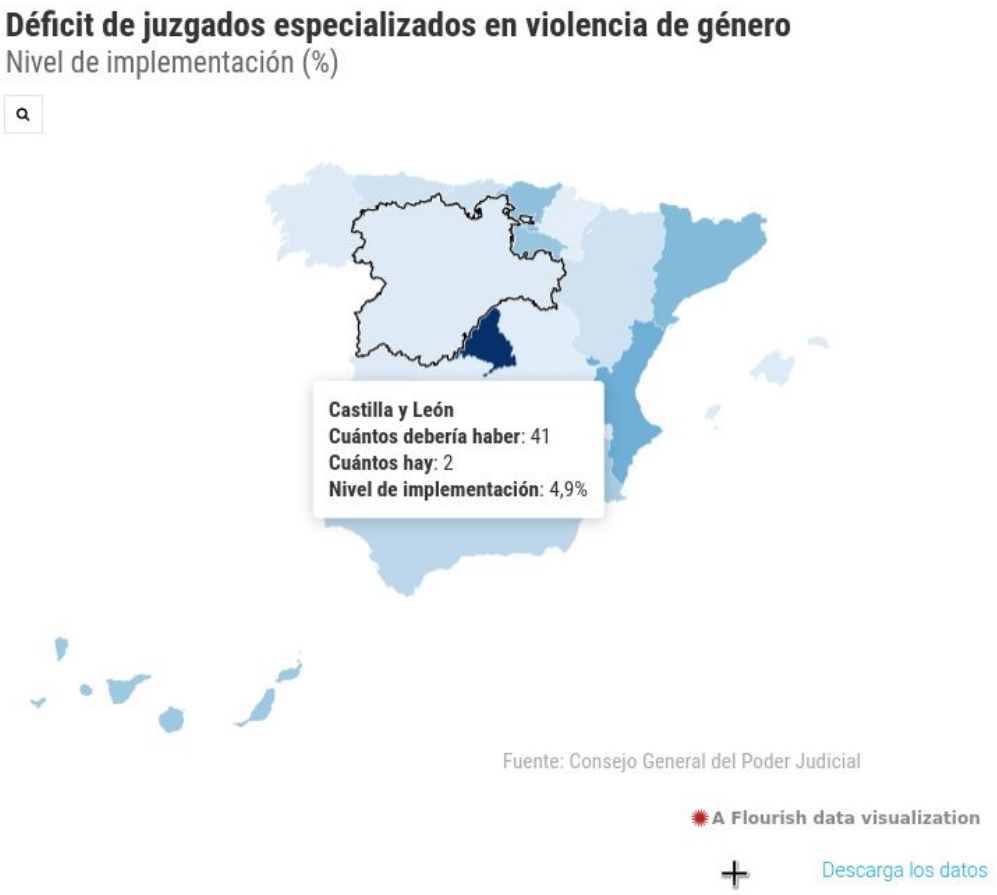
Figura 2.6 Comparación gráfica entre Canarias y la Comunidad Valenciana de 2008 a 2019



Fuente: Iniciativa Barcelona Open Data

Otro análisis a resaltar es el de los juzgados por comunidad autónoma y el cálculo del nivel de implementación en relación a la cantidad de entidades que deberían existir en cada comunidad (figura 2.7). A estos datos se podrían enlazar datos con partidos gobernantes o leyes autonómicas que contribuyen a la eliminación de la violencia de género.

Figura 2.7 Análisis por comunidad autónoma



Fuente: Iniciativa Barcelona Open Data

Estos proyectos son iniciativas para contribuir a la eliminación de la violencia contra las mujeres. La información es valiosa porque tiene un contexto, pero el problema es la limitación al tener una necesidad de generar análisis propios o con preguntas diferentes. Es decir, en estos proyectos la información está dada y los usuarios no pueden generar más, porque implicaría realizar un nuevo proceso de extracción de datos. Las gráficas son interactivas pero limitadas a la decisión de lo que se ha querido mostrar.

En un momento en el que el Big Data ha vuelto a tomar fuerza por las predicciones que pueden obtener, para Boyd y Crawford (2011, citado en Ardevól, 2017) “lo que es fundamental en Big Data son estas relaciones y los patrones que se pueden derivar de establecer conexiones entre distintos tipos y bases de datos”, por lo que aunque Big Data signifique grandes cantidades de datos, también es relevante el disponer relaciones entre estas distintas fuentes de datos, incluso datos de las mismas fuentes relacionando sus tablas porque permiten descubrir nuevos patrones o nuevas interpretaciones.

2.1. Propuesta

Tormenta tendrá una base de datos relacional donde almacenar datos abiertos de violencia de género del CGPJ, para que estos puedan ser gestionados y se puedan realizar análisis estadísticos profundos en tiempos más cortos que lo que tomaría un proceso ETL por completo. Para consultar esta base de datos y demostrar la potencialidad de la centralización de datos que ya han sido procesados, se utiliza una herramienta gráfica que permitirá ejecutar consultas SQL dentro de una plataforma web.

Los datos abiertos, al igual que los datos dentro de una empresa, necesitan ser tratados. De acuerdo con los creadores de Data Wrangler[9], una herramienta para la limpieza de datos elaborada por Stanford Vis Group[10], en sus conversaciones informales con analistas de datos han descubierto que el 80% del tiempo de un proyecto que finaliza en el análisis y visualización, se ocupa en la limpieza y estandarización de datos. Problemas como errores ortográficos o datos faltantes, duplicados son la parte más tediosa de estos proyectos. Recalcan que esto ocurre incluso con los grandes avances en tecnologías para la gestión y el análisis de datos (Kandel et al., 2011).

[9] Data Wrangler en su página web responde al problema del tiempo gastado en la manipulación de datos, literalmente dice: “Too much time is spent manipulating data just to get analysis and visualization tools to read it. Wrangler is designed to accelerate this process: spend less time fighting with your data and more time learning from it.”
Para utilizar este software no necesita una cuenta ni instalación, se utiliza en su dirección web: http://vis.stanford.edu/wrangler/. El proceso inicia con copiar y pegar datos en su formulario, después se puede acceder a las herramientas.

[10] The Stanford Visualization Group o Stanford Vis Group se encuentra disponible en: http://vis.stanford.edu/

Ya que la etapa de tratamiento de datos suele tomar gran cantidad de tiempo de un proyecto, se propone centralizar los datos en una base de datos relacional y ponerlos disponibles a través de una herramienta de consulta. De esta forma, los usuarios pueden generar nuevos análisis, establecer nuevos patrones o encontrar nuevas visualizaciones para su necesidad de comunicación.

Si bien Tormenta no es una solución original y novedosa, sino que combina soluciones ya existentes como la centralización de datos o las relaciones entre tablas para llegar a análisis más reflexivos, sí que implica un aporte en la reducción de tiempo que suponen los procesos con datos abiertos de violencia de género, una solución que aún no se ha encontrado implementada en proyectos a nivel nacional.

2.2. Marco referencial

En el marco referencial se describe el concepto desde el que se entiende la violencia de género para elaborar este trabajo, se explica el momento desde el que se empiezan a recabar datos desde los procesos judiciales y se da un contexto histórico del aparecimiento de la Ley Orgánica de Medidas de Protección Integral contra la Violencia de Género que hizo posible la recolección de datos de los procesos judiciales durante los 15 años que se cumplieron el 28 de diciembre pasado. También se habla del Observatorio, que es la institución que se encarga de abordar esta problemática desde la Administración de Justicia y publica los datos en el portal de estadísticas judiciales del CGPJ.

Para brindar una explicación detallada y mostrar la importancia de hacer un análisis temporal de los datos del CGPJ, se describe el proceso para recolectar los datos de los procesos judiciales en materia de Violencia de Género. Se debe tener en cuenta que los datos se recogen a partir de la Ley Orgánica de Medidas de Protección Integral contra la violencia de género, las entidades que han intervenido son el Observatorio de Violencia doméstica y de género desde la que se recopilan y analizan los datos; detrás del Observatorio están varias instituciones, entre ellas el CGPJ que tiene un portal en el que publica estadísticas judiciales.

2.2.1. Violencia de Género

La violencia de género se determina por un sistema patriarcal que privilegia lo masculino por sobre lo femenino y otras identidades. Ocurre en ámbitos distintos como

el doméstico o el laboral, en este último han aparecido términos como el techo de cristal. De acuerdo la Ley Orgánica de Medidas de Protección Integral contra la Violencia de Género:

> Art. 1 "La presente Ley tiene por objeto actuar contra la violencia que, como manifestación de la discriminación, la situación de desigualdad y las relaciones de poder de los hombres sobre las mujeres, se ejerce sobre éstas por parte de quienes sean o hayan sido sus cónyuges o de quienes estén o hayan estado ligados a ellas por relaciones similares de afectividad, aun sin convivencia."
> (BOE núm. 323, 2004, p. 10)

Durante el primer encuentro feminista en América Latina y el Caribe celebrado en Bogotá en julio de 1981, (Varela, 2014) describe la violencia de género como "una realidad sistemática que abarcaba desde agresiones domésticas a violaciones, desde tortura sexual a violencia de estado, incluyendo abusos a mujeres prisioneras políticas". El 25 de noviembre de 1999 es el día que se recuerda como el Día Internacional contra la violencia hacia las mujeres, aunque desde organizaciones feministas se pidió que existiera mucho antes.

2.2.2. Ley Orgánica de Medidas de Protección Integral contra la Violencia de Género

Esta ley fue promulgada el 29 de diciembre de 2004 gracias a la lucha de organizaciones de mujeres, según se reconoce en la actual sección sobre la exposición de los motivos en BOE núm. 323 (2004), organizaciones para erradicar todas las formas de violencia de género. De hecho, Varela (2004) explica que el trabajo detrás de esta publicación empezó en 1998 desde las organizaciones de mujeres feministas que trabajaban en el estudio de la violencia de género y la atención a las víctimas. Tenían varios objetivos, y uno de ellos era recolectar datos. Tras 6 años de lucha, en 2004 se aprobó por unanimidad en el Congreso de los Diputados bajo el gobierno de José María Aznar .

2.2.3. Observatorio de Violencia doméstica y de género

Antes de la promulgación de la Ley Orgánica de Medidas de Protección Integral contra la Violencia de Género, el CGPJ ya contaba con el Observatorio de Violencia doméstica y de género. El objetivo del Observatorio ha sido desde entonces abordar el tratamiento de la violencia doméstica y de género desde la Administración de Justicia,

por lo que se encarga de recopilar y analizar los datos obtenidos de las estadísticas judiciales.

El observatorio fue creado el 26 de septiembre de 2002. Actualmente continúa y tiene su sede en Madrid. Además del CGPJ, las instituciones que están detrás son: el Ministerio de Justicia, el Ministerio de Sanidad, Servicios Sociales e Igualdad, el Ministerio del Interior, la Fiscalía General del Estado, las CCAA con competencias transferidas en Justicia, el Consejo General de la Abogacía Española y el Consejo General de Procuradores de España (CGPJ, n.d.).

2.2.4. Datos del CGPJ

El CGPJ inició la recolección de datos estadísticos unos meses antes de la Ley anteriormente descrita. Las descripciones con las que se recogían los datos se ampliaron desde el 29 de junio del mismo año al crearse los juzgados de violencia sobre la mujer y las secciones especializadas dentro de las Audiencias Provinciales de toda España. A partir de su creación, los campos para recolectar información son actualizados con frecuencia. En el portal, el CGPJ aclara que "se han introducido nuevos datos que permiten aproximarse a aspectos que se han considerado precisados de medición". Los datos y los informes son publicados en el portal estadístico, y según el calendario preestablecido disponible en el mismo, hay una publicación trimestral y la publicación anual puede darse hasta 4 meses después del año finalizado.

A partir de la creación de los juzgados[11] y la publicación de la Ley se elaboraron boletines exclusivos, con periodicidad trimestral y que debían ser remitidos por "todos los Juzgados de Violencia sobre la Mujer, con competencias exclusivas y con competencias compartidas al Consejo General del Poder Judicial para su tratamiento estadístico" (CGPJ, n.d.). Es con esta información con la que el CGPJ realiza boletines estadísticos, estudios e informes a través de los cuales pone a disposición los datos de la Administración de Justicia.

[11] En 2015, la entonces presidenta del Observatorio de Violencia doméstica y de género, Ángeles Carmona Vergara, afirmó que eran 106 juzgados exclusivos y 355 compatibles en toda España.

De acuerdo con la información del portal estadístico[12] del CGPJ acerca de los datos de violencia de género, estos datos contienen “datos recogidos en los boletines estadísticos trimestrales de los juzgados de violencia contra la mujer, y los apartados de violencia contra la mujer de los boletines de los juzgados de instrucción y primera instancia e instrucción, de lo penal, de menores y audiencias provinciales. Se ofrecen resultados a nivel de tribunal superior y justica, de provincia y de partido judicial (solo para los datos de los juzgados de violencia contra la mujer)”

[12] Para acceder a la información del portal estadístico con datos sobre violencia sobre la mujer se puede acceder a: https://cutt.ly/fzRBeon

3.Análisis del problema

El problema para tratar los datos de los órganos judiciales, como se aclara en el informe de 2015 de la Ley Orgánica de Medidas de Protección Integral contra la Violencia de Género, es que "desgraciadamente siguen existiendo carencias, usándose conceptos y definiciones distintas, lo que redunda en la dificultad en tener una visión completa y exacta de la violencia contra la mujer." (CGPJ, 2015, p.1). Estas carencias y distintas definiciones se ven reflejadas en los datos que publican, lo que hace evidente la necesidad de tratar los datos antes de utilizarlos.

3.1. Documentación de posibles herramientas

En la búsqueda de disminuir el tiempo que tarda la limpieza de datos se realiza una documentación de herramientas web para escoger la más conveniente. Aunque finalmente se escoge la creación de código propio, en la siguiente sección se explica la historia y usos de las herramientas han sido tomadas, en su mayoría, de las recomendaciones de SocialTIC y Escuela de Datos[13], porque son asociaciones que apoyan a otras comunidades en la creación de proyectos con datos. Al inicio de cada apartado se nombra la fase del proceso en la que pueden ser utilizadas, de acuerdo con el proceso ETL.

Extracción de datos

PDF Tables, anteriormente llamada ScraperWiki y con más funcionalidades[14], creada por la empresa The Sensible Code Company, se encarga de convertir archivos en formato PDF a Excel, CSV, XML o HTML**.** Por ahora gratuita, para las primeras 25 páginas no se necesita crearse una cuenta, solo debe cargarse el archivo, de forma inmediata se obtiene una vista de su conversión a tablas. Hasta las

[13] Escuela de Datos se define como una comunidad hispanohablante compuesta por una red de activistas y personas relacionadas con la comunicación que buscan aprovechar los datos en el cambio social. Disponibles en: https://escueladedatos.online/

[14] ScraperWiki era una herramienta gratuita para *scraping*. La nueva herramienta de la compañía se llama QuickCode y además de las anteriores funcionalidades describe la creación de un entorno con Python y R dedicado a Economistas, personas de estadística y Data Managers que han iniciado en la programación.

siguientes 50 páginas se necesita una cuenta y después de esta cantidad se requiere conexión a una clave de API única sin costo. La herramienta detecta las tablas y las recrea, con el resto del texto elabora otras tablas donde introduce la información indistintamente. En la explicación, la compañía creadora explica que es capaz de convertir 3 páginas por segundo[15]; también asegura que cada mes tiene, al menos, 50 mil nuevos visitantes.

Transformación y limpieza de datos

Open Refine, inicialmente conocida como Freebase Gridworks, pasó a llamarse Google Refine cuando Google compró la empresa, Metaweb, y en 2012 anunció que dejaría de darle soporte. Desde ese momento pasó a ser parte de la comunidad Open Source. Esta es una aplicación de escritorio que se utiliza para limpiar y transformar datos. A esta aplicación se puede acceder a través de su instalación local disponible para Mac, Windows y Linux. Para acceder a los datos se abre la aplicación local en el navegador y para empezar se carga el fichero de datos a limpiar. El uso es muy parecido a manejar un archivo en Excel. Para eliminar datos que están acompañando, por ejemplo a los años, se puede indicar el tipo de carácter que se desea eliminar y darle click a aceptar. La plataforma cuenta con tutoriales, en inglés y se han encontrado otros externos en español.

Carga de datos

Socrata es un software desarrollado por la empresa Tyler Technologies de Texas, Estados Unidos. Este es un sistema de gestión de datos abiertos de pago que ha sido desarrollado parcialmente en Open Source por lo que está en dependencia del fabricante y no de la comunidad.

DataHub es un término que además de referenciar a sistemas de gestión de grandes volúmenes de datos, es el nombre de una plataforma que permite publicar y desplegar datos ya sea para personas, organizaciones o equipos. Este software necesita instalación local a través de la terminal y la creación de una cuenta en su

[15] Acerca de las dos herramientas de The Sensible Code Company: https://sensiblecode.io/

portal web para desplegar los datos. La creación de la cuenta otorga los permisos[16] estándar de la suscripción *Free*: permite el almacenamiento de hasta 1GB con una banda ancha del mismo valor. Sin embargo, solo con la suscripción *premium*, que no tiene un precio sino un proceso de contacto con sus administradores, se tiene acceso 10GB, a actualizaciones garantizadas e integraciones de flujos de trabajo con paquetes de Python y NPM.

CKAN es una plataforma desarrollada por la OKFN y los creadores de DataHub[17]. Su desarrollo ha sido direccionado para ser código abierto y software libre, por lo que tiene una comunidad fuerte detrás. Es sobre todo utilizada para crear portales de datos abiertos. Cuenta con algunas referencias mundiales como data.gov, data.gov.uk, Berlin Open Data, canada.ca (Pollock & Kariv, n.d.). En su página web[18] se define como una herramienta similar a WordPress por su facilidad para manejar y publicar datos.

[16] Los planes de publicación y permisos por tipos de cuentas están disponibles en: https://datahub.io/pricing

[17] CKAN es una plataforma que se encuentra entre los productos desarrollados por esta empresa. La información está disponible en: https://datahub.io/hire-us

[18] Acerca de CKAN: https://ckan.org/about/

4.Solución propuesta

A continuación se describe la solución propuesta, el plan de trabajo, se detalla la creación de la secuencia de comandos, la utilización de la herramienta Redash, la interacción de los archivos que se ejecutan en cada fase del proceso ETL; se comenta el código y se explica cómo ejecutarlo. Además, se presenta la arquitectura creada con Docker Compose y la utilidad que brinda.

Tormenta se crea como una herramienta de software libre, por lo que podrá ser usada, estudiada, distribuida y mejorada. De ahí que las librerías o herramientas que utiliza son lo más accesibles posible. Varias de las herramientas descritas anteriormente eran de uso gratuito pero han reducido su funcionalidad para la versión no *premium*. Por esto y por la capacidad de adaptación a los requerimientos, se escoge elaborar la secuencia de comandos con JavaScript para realizar el proceso de ETL.

Un ejemplo de las herramientas que han cambiado sus políticas de uso es PDF Tables, anteriormente llamada ScraperWiki, ha limitado el uso gratuito para escanear tablas en PDFs. En su página los planes de pago inician con 1000 páginas[19], las cuales vencen si no han sido utilizadas durante ese año. CKAN es la opción más próxima a las necesidades del proyecto porque es una herramienta de código abierto, pero se ha descartado porque no es un software destinado a establecer relaciones entre las distintas fuentes de datos. La solución escogida responde a la disponibilidad con la que se puede reutilizar, extender, estudiar y modificar la herramienta, pues se considera que los beneficios sociales son más impactantes con herramientas de Software Libre.

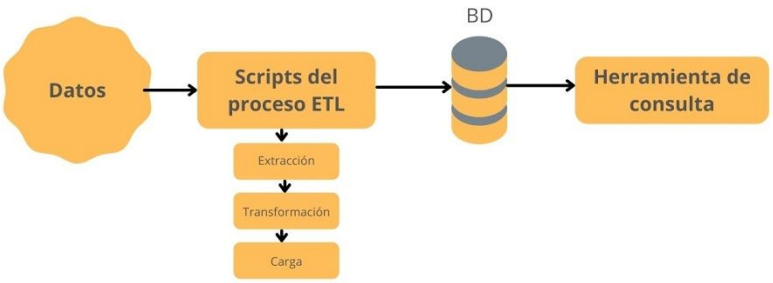
4.1. Plan de trabajo

El tratamiento de datos a través del proceso ETL se divide en 3 fases: extracción de datos de las fuentes escogidas, limpieza de datos a través de la selección y criterios de normalización, y la carga de datos en una BD. Tras la primera carga de datos, estos se revisan de forma manual teniendo en cuenta los criterios de normalización, si hay algún error se corrige y se vuelven a cargar los datos. Cuando los datos están

[19] Esta es la sección de precios de PDFTables: https://pdftables.com/pricing. Último acceso: 02-12-2020

listos se utiliza la herramienta llamada Redash, ésta se conecta a la BD de Tormenta y permite realizar consultas SQL. La primera prueba de Tormenta es el análisis estadístico disponible en el apartado de pruebas de este trabajo.

Figura 4.1 Esquema de plan de trabajo



Fuente: elaboración propia

4.2. Diseño de la solución

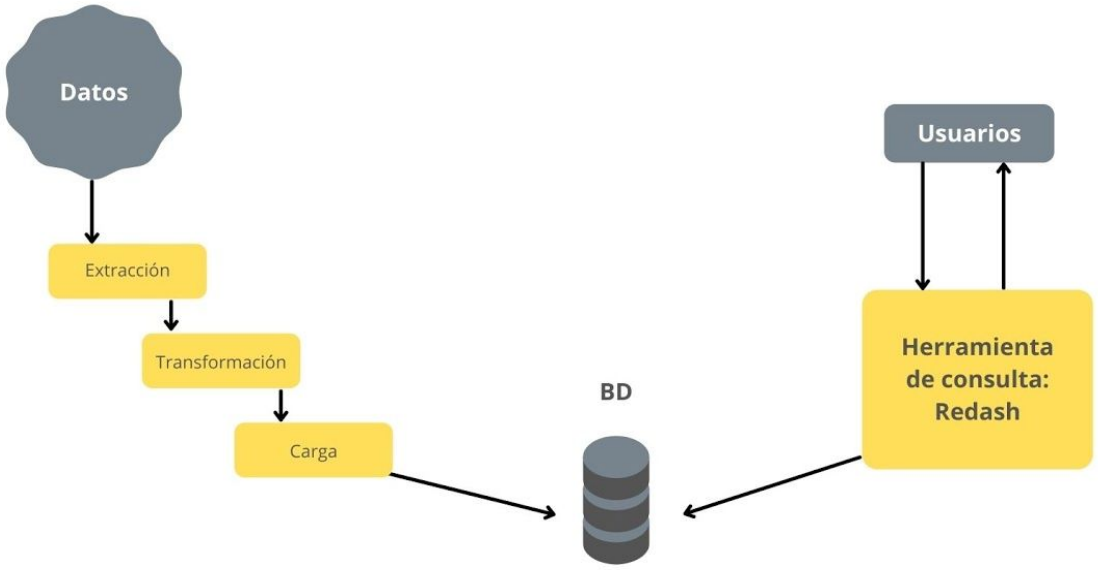
Los datos tratados se disponen para su consulta o descarga, a través de Redash, esta es una herramienta de Software Libre que permite conectar, manipular, compartir y visualizar varias fuentes de datos para su consulta. Existen diferentes software con funcionalidades de consulta como las de esta herramienta, por ejemplo Tableau o Google Data Studio, pero Redash tiene las siguientes ventajas: libre, sin coste y la posibilidad de su instalación en un entorno local. Para este proyecto, Redash funciona como una caja negra que se utiliza y no se indaga desde el desarrollo de software. Al realizar su configuración se pueden conectar múltiples tipos de BD, siempre que sean parte de las soportadas. El volumen de datos no es afectado por el tipo de datos a utilizar, para Redash y para Tormenta el único requisito es que sea una BD relacional como MariaDB.

4.3.Arquitectura del sistema

El punto de partida de la arquitectura de las secuencias de comandos es la extracción de los datos desde sus fuentes. Dependiendo del formato con el que se obtienen, se toman decisiones para realizar su transformación. Los datos se encuentran en los formatos XLS, en el caso del CGPJ y en CSV los del INE. Este último formato también se utilizó para relacionar la provincia con su comunidad autónoma. De ahí que para el proceso de transformación se soporte archivos de formatos XLS y CSV. Después, estos datos se convierten en un objeto de JavaScript para poder trabajarlos

directamente con el lenguaje de programación. Aquí entra la etapa de elección de hojas de trabajo y de normalización en la que se registran las provincias, las comunidades autónomas y las variaciones al momento de nombrarlas. A continuación se cargan los datos en una BD relacional. Se hace la revisión de los datos, si los procesos anteriores han ido bien se inicia la conexión con la herramienta de consulta. A través de esta herramienta, los datos se ponen disponibles para los usuarios. En la siguiente figura se muestra la relación entre componentes antes explicados:

Figura 4.3.1 Relación entre los componentes de Tormenta



Fuente: elaboración propia

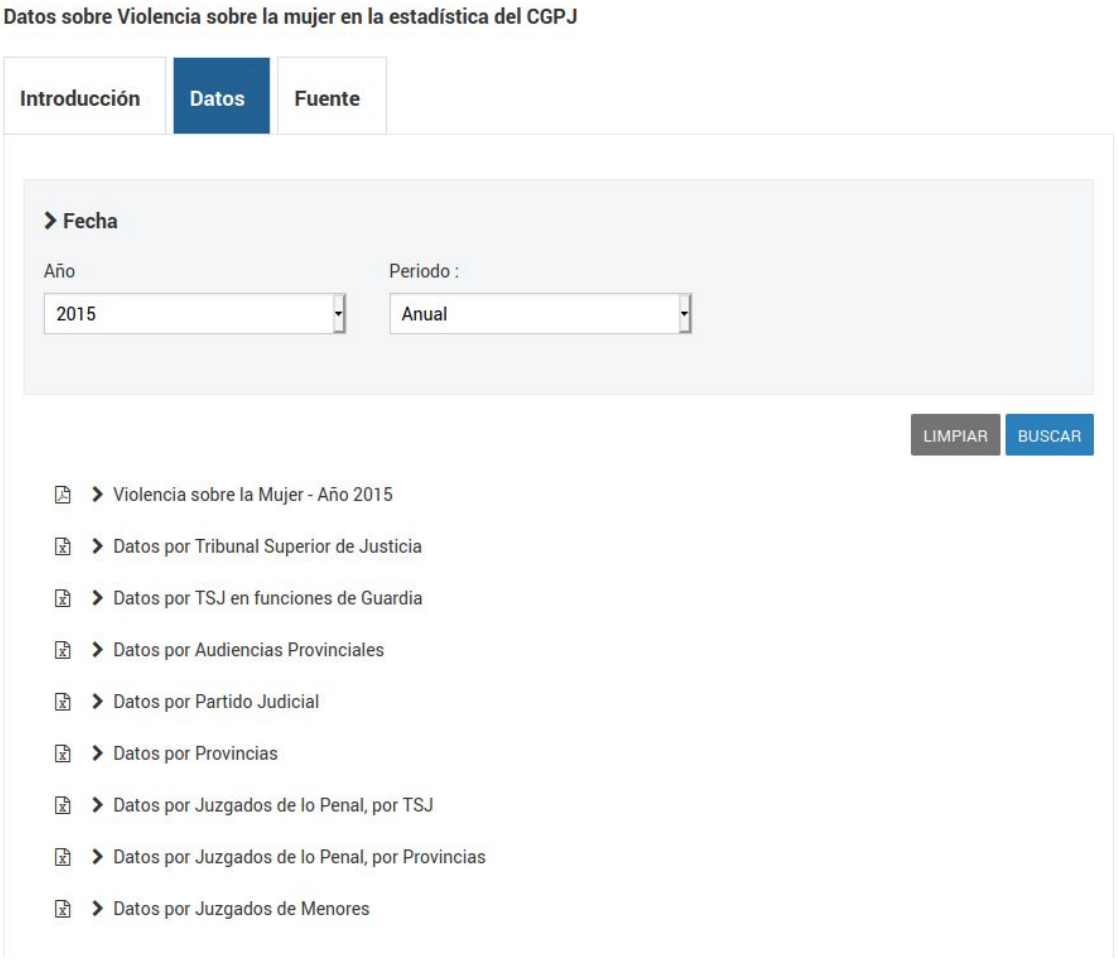
4.3.1. Proceso ETL

Extracción

Para el primer proceso: la extracción de datos, se ha tomado en cuenta que en España existen varias fuentes con datos de los órganos judiciales: Estadísticas de violencia de género del Ministerio de Igualdad o BBDD Estadística Judicial. Para este trabajo se han tomado los datos de la fuente principal, que es el portal estadístico del CGPJ. En este portal se aclara que los datos se pueden publicar hasta 3 meses después del año finalizado, por lo que 2020 no ha podido ser añadido. Por otro lado, la fuente de datos de la población es el INE, mientras que la información de la relación entre provincias y comunidades autónomas es de conocimiento público, así que para elaborarla no ha hecho falta su extracción.

Uno de los problemas para la extracción de datos del CGPJ, es que no se pueden realizar peticiones a servicios web. Se debe realizar una búsqueda manual en la que se especifica el año y el periodo, como se muestra en la siguiente figura. Para realizar la extracción automática se guarda la dirección web del archivo “Datos por Provincias”. El proceso debe realizarse por año y por periodo anual.

Figura 4.3.1.1 Datos inexistentes de 2020 en el CGPJ



Fuente: Portal estadístico del CGPJ, sección de Datos penales, civiles y laborales

Para los datos poblacionales del INE también se buscaron servicios web, pero hasta la fecha de la elaboración de este trabajo no se encontraron. La mejor opción fue realizar una descarga manual de los datos del total de la población. Los datos a obtener se pueden seleccionar, tal como se muestra en la siguiente figura, se puede escoger una opción o todas, por ejemplo en la primera columna se escoge por total y/o hombres y/o mujeres, y se han seleccionado hombres y mujeres.

Figura 4.3.1.2 Plataforma del INE para descargar datos de Población residente por fecha, sexo y edad

Fuente: INE

Esta fuente permite consumir datos de diferentes formatos: XLS, XLSx, CSV separado por (;) o por tabuladores, Pc-Axis, Json, texto plano: separado por tabuladores, (,) o (;). Mientras se realizaba el proceso de extracción de este trabajo, la descarga de JSON daba problemas por exceso de información, así que se escogió el formato CSV que de acuerdo con los costos y beneficios definidos por Tim Berners-Lee, antes mencionados, tiene 3 de 5 estrellas.

Transformación

En esta etapa, de acuerdo con Gerardo (2008, p. 21) se incluye la "limpieza de datos, integración de formato, integración semántica, conversión de estructuras internas, integración de datos, resumen o agregación de datos". La integración semántica no se realizó por la falta de conocimientos jurídicos. La limpieza de datos incluyó la normalización de las provincias, por ejemplo, para Valencia se encontraron las variaciones: VALÈNCIA/VALENCIA y VALENCIA/VALÈNCIA. Esta normalización se enlista en los anexos. También se encontraron cambios en los nombres de las hojas internas de los archivos y aunque estos nombres no influyen en la BD de Tormenta, se

recogen estos cambios en la siguiente tabla. Por otro lado, a partir de 2016 se aumentan 3 columnas de datos que se definieron como nulos para 2015.

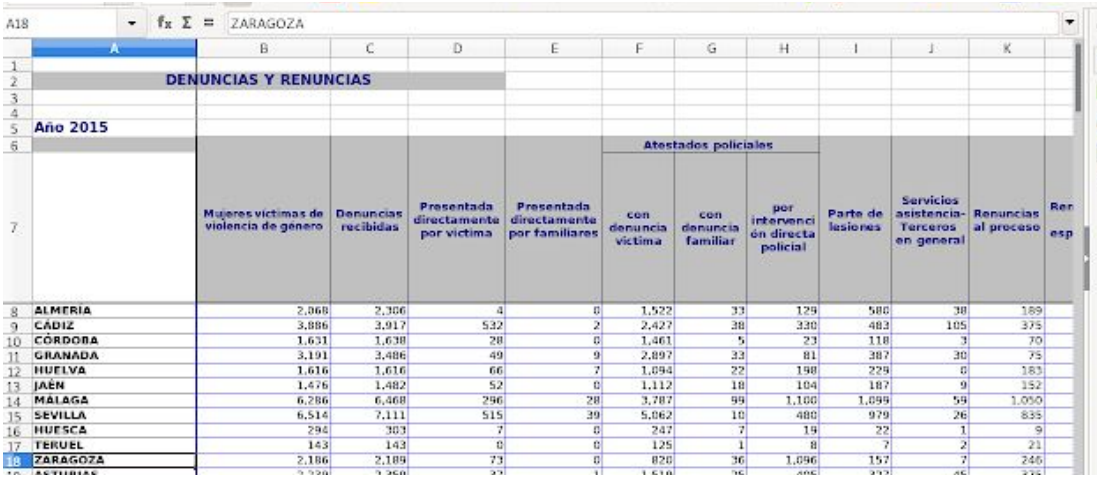
Tabla 4.3.1.1 Registro de cambios necesarios en la transformación de datos del CGPJ

Año	Hoja	Observación
2016	“Delitos”	Desde este año se incrementaron las columnas: 1. Contra la intimidad 2. Derecho a la propia imagen 3. Contra el Honor
2016	“Órdenes y Medidas”	En el año anterior la hoja de trabajo se llama: “Órdenes” en 2016 se cambia por “Órdenes y medidas”. Se mantienen los mismos campos de datos del año anterior.
2018	“Relación Víctima_Denunciado”	En los años anteriores esta hoja de trabajo se llamaba “Relación”.

Fuente: Elaboración propia

Debido al límite de tiempo por la realización del proceso de transformación de datos con una secuencia de comandos propia y la necesidad de probar la solución, para realizar este trabajo, se escogieron las hojas de: “Delitos ingresados”, “Órdenes de protección: sexo y nacionalidad”, “Relación”, “Denuncias y renuncias”. Los problemas fueron sobre todo en el cambio de formato desde 2018. La librería utilizada para convertir el XLS a un Objeto de JavaScript, llamada “node-xlsx”, ignora todas las filas que están vacías al inicio o al lateral. En el año 2015 (figura 4.3.1.3) la librería ignora la primera fila y toma las siguientes como válidas mientras que en el archivo de 2018 (figura 4.3.1.4) ignoraba la columna A y hasta la fila 9, desde la que empieza a contar. La solución fue definir, de cada archivo, la fila por la que se esperaba que se empiecen a guardar los datos.

Figura 4.3.1.3 Archivo de Excel con datos de 2015



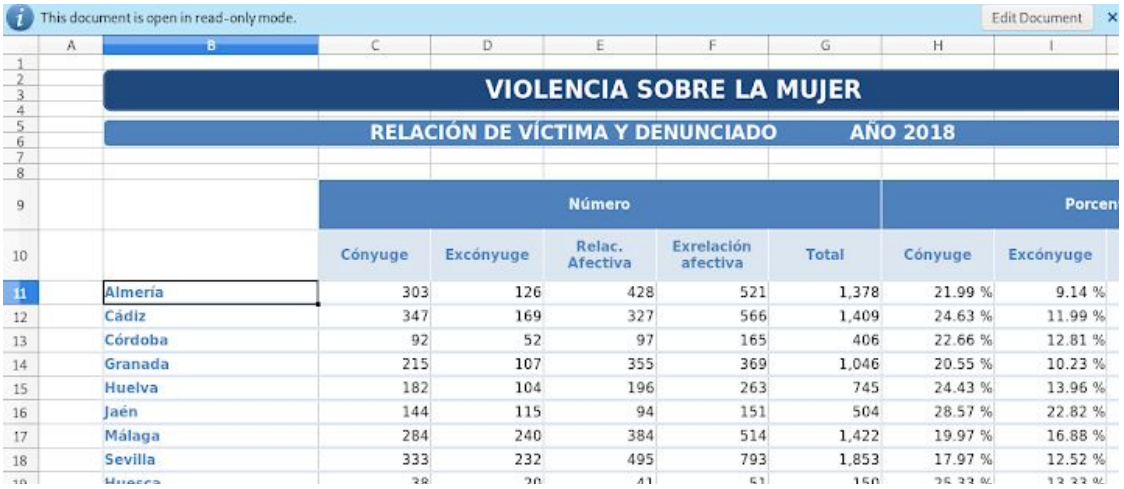
DENUNCIAS Y RENUNCIAS

Año 2015

	Mujeres víctimas de violencia de género	Denuncias recibidas	Presentada directamente por víctima	Presentada directamente por familiares	Atestados policiales: con denuncia víctima	Atestados policiales: con denuncia familiar	Atestados policiales: por intervención directa policial	Parte de lesiones	Servicios asistencia-Terceros en general	Renuncias al proceso	Ren esp
ALMERÍA	2,868	2,306	4	0	1,522	33	129	580	38	189	
CÁDIZ	3,886	3,917	532	2	2,427	38	330	483	105	375	
CÓRDOBA	1,631	1,638	28	0	1,461	5	23	118	3	70	
GRANADA	3,191	3,486	49	9	2,897	33	81	387	30	75	
HUELVA	1,616	1,616	66	7	1,094	22	198	229	0	183	
JAÉN	1,476	1,482	52	0	1,112	18	104	187	9	152	
MÁLAGA	6,286	6,468	296	28	3,787	99	1,100	1,099	59	1,050	
SEVILLA	6,514	7,111	515	39	5,062	10	480	979	26	835	
HUESCA	294	303	7	0	247	7	19	22	1	9	
TERUEL	143	143	0	0	125	1	8	7	2	21	
ZARAGOZA	2,186	2,189	73	0	820	36	1,096	157	7	246	

Fuente: CGPJ (2015)

Figura 4.3.1.4 Archivo de Excel con datos de 2018



VIOLENCIA SOBRE LA MUJER

RELACIÓN DE VÍCTIMA Y DENUNCIADO AÑO 2018

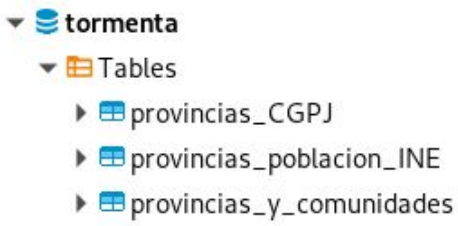
	Número					Porcen	
	Cónyuge	Excónyuge	Relac. Afectiva	Exrelación afectiva	Total	Cónyuge	Excónyuge
Almería	303	126	428	521	1,378	21.99 %	9.14 %
Cádiz	347	169	327	566	1,409	24.63 %	11.99 %
Córdoba	92	52	97	165	406	22.66 %	12.81 %
Granada	215	107	355	369	1,046	20.55 %	10.23 %
Huelva	182	104	196	263	745	24.43 %	13.96 %
Jaén	144	115	94	151	504	28.57 %	22.82 %
Málaga	284	240	384	514	1,422	19.97 %	16.88 %
Sevilla	333	232	495	793	1,853	17.97 %	12.52 %

Fuente: CGPJ (2018)

Carga

Los datos tratados se cargaron en el software de gestión de datos, MariaDB, a través de sentencias de SQL. Este proceso se realizó varias veces, pues dependía de que el proceso anterior haya cumplido con todos los cambios. Las tablas con las que cuenta Tormenta son las descritas en la siguiente figura:

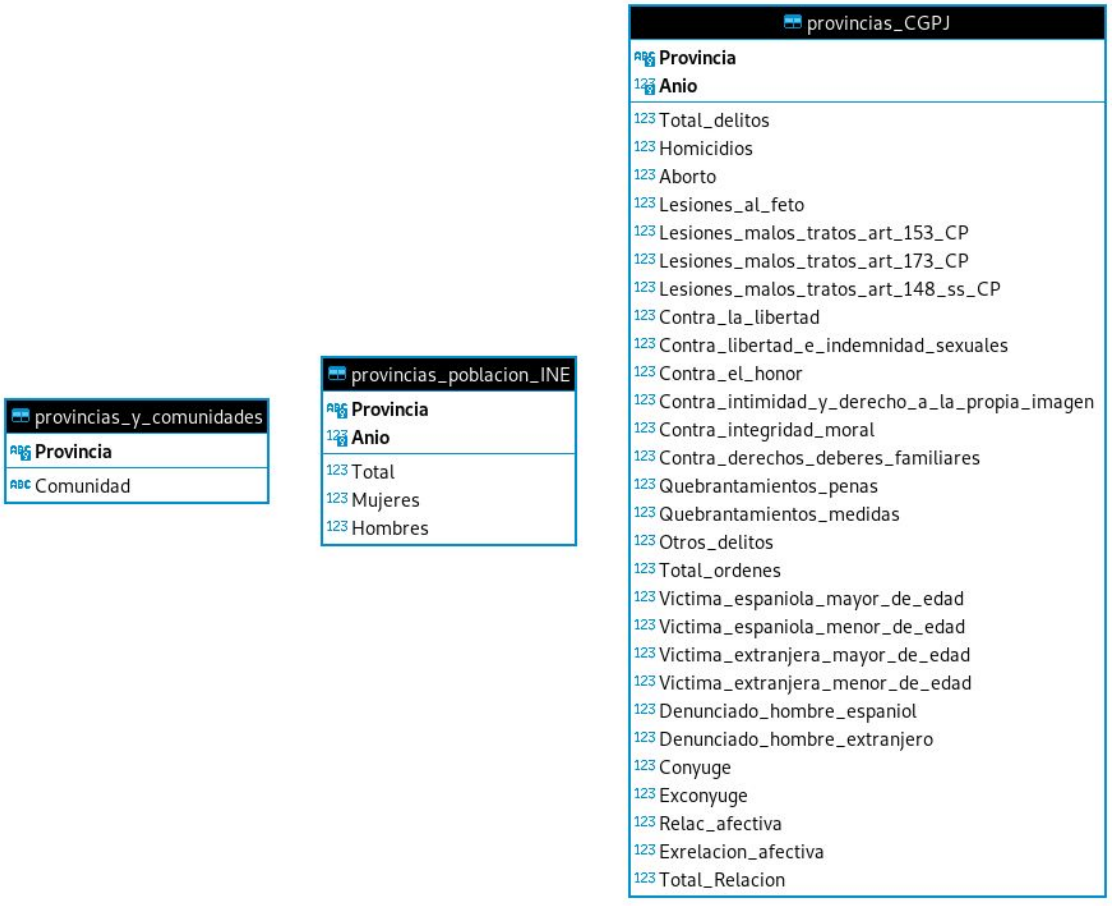
Figura 4.3.1.5 Tablas cargadas en la BD de Tormenta



Fuente: Software DBeaver

Las tablas o entidades conservan sus nombres de acuerdo a la fuente de datos de las que provienen, por ejemplo “provincias_CGPJ” contiene los datos del CGPJ. En la siguiente figura se muestran los datos que contiene cada tabla:

Figura 4.3.1.6 Digrama del contenido de cada tabla



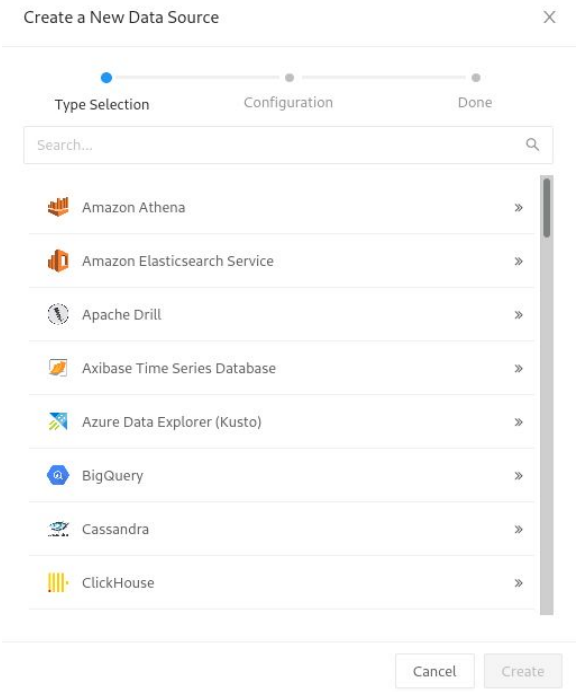
Fuente: Software DBeaver

Hasta aquí se explica el proceso ETL y la elaboración de secuencia de comandos para realizarlo. A continuación se explica la conexión con la herramienta Redash para la consulta de estos datos ya tratados.

4.3.2. Conexión con la herramienta de consulta

Redash es un software que está preparado para establecer conexiones con multitud de tipos de bases de datos diferentes, en la siguiente figura se pueden apreciar solo algunas de ellas: Amazon Atena, BigQuery, Casandra, Click House. La siguiente figura también es una muestra de los 3 pasos para añadir una nueva BD. Después de elegir el motor de BD se define la configuración, con los parámetros que se indicarán más adelante, y finalmente se testea la conexión entre Redash y la BD.

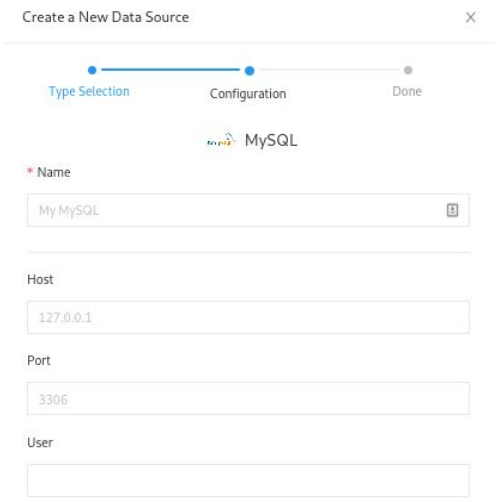
Figura 4.3.2.1 Muestra del listado de BD que se pueden configurar en Redash



Fuente: Redash

Para establecer la conexión, se deben escribir los parámetros de conexión obligatorios: usuario, contraseña, nombre de la base de datos, dirección y puerto. Algunos de esos parámetros se muestran en la siguiente figura. En el caso de Tormenta se ha configurado una BD con MySQL.

Figura 4.3.2.2 Campos de configuración de Redash para añadir una nueva BD



Fuente: Redash

4.4. Diseño detallado

El código de las secuencias de comandos está disponible en el repositorio de GitLab en la dirección web: https://gitlab.com/annchoa/tormenta. Este código se crea con Docker Compose, para facilitar el desarrollo y la reutilización del código, la única dependencia e instalación que se necesita en un entorno local, para su réplica, es Docker y Docker Compose. En la siguiente figura se ve la configuración del archivo de Docker Compose dividido por: la versión, los servicios y la red para comunicar estos.

Figura 4.4.1 Versión y servicios de Docker Compose

```
version: "2.2"
services:
  db:
    image: mariadb:latest
    environment:
      MYSQL_ROOT_PASSWORD: root
      MYSQL_DATABASE: tormenta
    ports:
      - 3306:3306
    networks:
      - tormenta
  etl:
    build: ETL/.
    volumes:
      - "./ETL:/app"
    networks:
      - tormenta

networks:
  tormenta:
    name: tormenta
```

Fuente: Código del proceso ETL de Tormenta

A continuación se muestran los comandos para iniciar el proceso, ya que se ha desarrollado una secuencia de comandos, estos deben ser ejecutados a través de una Terminal. El primer comando está definido en la siguiente tabla: `docker-compose up`. Este comando levanta los servicios definidos en el Docker Compose (figura 4.4.1). Los siguientes comandos, definidos de acuerdo a las fases del proceso, se pueden ejecutar por separado, teniendo en cuenta que la transformación y la carga comparten los procesos porque no se han desacoplado para que se puedan ejecutar por separado.

Tabla 4.4.1 Versión y servicios de Docker Compose

Proceso	Comando	Observaciones
Iniciar servicios de Docker Compose	`docker-compose up`	Este comando levanta los contenedores para que se puedan ejecutar los comandos. Es esencial para iniciar cualquier fase.
Etapa de **Extracción** del proceso ETL	`docker-compose exec etl npm run extract`	Para ejecutar el proceso de extracción.
Etapa de **Extracción** del proceso ETL	`docker-compose exec etl npm run transformAndLoad`	Para ejecutar los procesos de transformación y carga de datos.
Levantar servicio de BD	`docker-compose exec db mysql -u root -proot tormenta`	Los datos limpios, que son objetos de JavaScript, se pueden cargar en la BD tras acceder al servicio de la BD con este comando

Fuente: elaboración propia

Todos los comandos tienen como punto de entrada el archivo *cli.js* (figura 4.4.2). Aquí se extrae el subcomando de los argumentos y se invoca a la función `execute` del archivo *commands.js.*

Figura 4.4.2 Script del archivo cli.js

Fuente: Código del proceso ETL de Tormenta

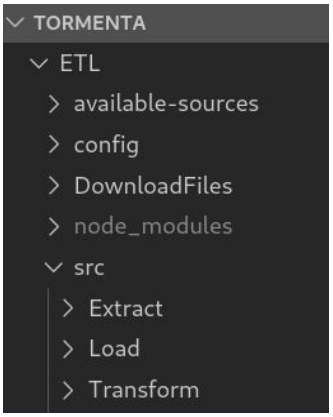
El archivo commands.js (figura 4.4.3) cuenta con un listado de comandos habilitados que van asociados a una función por cada comando, esta función es llamada desde la función `execute`. Estas funciones, a su vez, llaman a las clases de cada una de las fases, en este caso de la extracción. Por ejemplo, si se ejecuta el comando `docker-compose exec etl npm run extract`, el último parámetro, extract llamará a la función `extractAll`, tal como se muestra en la línea 7.

Figura 4.4.3 Script del archivo cli.js

Fuente: Código del proceso ETL de Tormenta

Cada fase se encuentra en una subcarpeta cuyo nombre corresponde a la etapa del proceso. En la siguiente figura se puede observar que dentro de la carpeta *src* se muestran, por orden alfabético, las carpetas *Extract, Load y Transform.*

Figura 4.4.4 División de fases del proceso por carpetas



Fuente: Código del proceso ETL de Tormenta

Las fuentes para extraer los datos y las necesidades de normalización se configuran en un archivo en formato JSON que se debe colocar en la carpeta *config*. Esta carpeta no se sube al repositorio de GitLab para que quien se clone este proyecto pueda cargar los archivos que quiere tratar. La lista de fuentes extraíbles soportadas están disponibles para su consulta en la carpeta *available-sources*, estas fuentes pueden servir de ejemplo para definir nuevas o para ser utilizadas. Hasta el momento se da soporte a archivos de formato XLS y CSV. La configuración del archivo JSON se crea para que sea reutilizada por usuarios con pocos conocimientos de programación.

En la siguiente figura se ve un ejemplo del archivo *Provincias_CGPJ_2015*, aquí se definen las características del proceso ETL en el siguiente orden: el nombre que se le dará al archivo descargado, la URL de descarga, el nombre de la tabla en la que se cargarán los datos procesados, el tipo de fuente extraída, los campos que se necesitará extraer o las transformaciones que se efectúan sobre los datos.

Figura 4.4.5 Ejemplo de una fuente soportada para el proceso ETL

Fuente: Código del proceso ETL de Tormenta

Ahora bien, en la subcarpeta *Extract* se encuentra una clase que corresponde al único método de extracción que se soporta actualmente: la descarga mediante una petición HTTP, como se muestra en la siguiente figura, entre las importaciones necesarias solo existe esta petición. En el futuro es posible que soporten otros métodos de extracción como FTP o peticiones a servicios web.

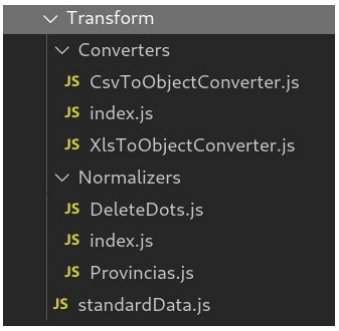
Figura 4.4.6 Importaciones necesarias en el archivo para extraer datos

Fuente: Código del proceso ETL de Tormenta

En la subcarpeta *Transform* se encuentran dos carpetas. La primera es *Converters*, cuya clase sirve para transformar el formato descargado a un formato común de trabajo. Por ejemplo de XLS o CSV a objetos de JavaScript, que será el formato utilizado en el resto del proceso. Por tanto, cada formato de origen tiene una `class Converter` asociada. En la siguiente figura se puede observar XlsToObjectConverter.js y CsvToObjectConverter.js. La carpeta *Normalizers* contiene archivos para aplicar transformaciones que normalizan campos específicos, por

ejemplo: la normalización del nombre de provincias, cuyo archivo tiene el nombre *Provincias.js*, o el formato adecuado para los números, *DeleteDots.js*.

Figura 4.4.7 Estructura de carpetas del proceso de transformación



Fuente: Código del proceso ETL de Tormenta

Antes de realizar la carga de los datos procesados en la BD se necesitan crear las tablas en el gestor de datos para que estos puedan ser incrustados. Después de hacerlo inicia la última fase, la carga, dentro de la carpeta *Load* se encuentra un único archivo, ya que actualmente la carga solo se realiza en un único medio. Los datos están estandarizados en objetos de JavaScript, por lo que no se precisan clases extra para mapear diferentes orígenes a un mismo o diferentes destinos. El archivo mencionado se llama *SqlLoad.js (figura 4.4.8)* e incluye una clase que genera una consulta de inserción con los datos recogidos dentro de una BD relacional, MariaDB.

Figura 4.4.8 Construcción de la consulta SQL para la inserción en la BD

SqlLoad.js

ETL > src > Load > SqlLoad.js > SqlLoad > getHeaders

```
1   import mysql from 'mysql'
2   import normalizers from '../Transform/Normalizers/index.js'
3
4   export default class SqlLoad {
5       constructor(config) {
6           this.config = config
7       }
8
9       buildQuery(entries) {
10          const columns = this.getHeaders()
11          const rows = this.getValues(entries)
12          const table = this.config.table_name
13
14          const queryString = `INSERT INTO ${table} (${columns}) VALUES ${rows}`
15          //console.log(queryString)
16          this.executeQuery(queryString)
17      }
```

Fuente: Código del proceso ETL de Tormenta

En este punto se realiza la conexión con Redash. Esta herramienta está configurada para conectarse a la misma red de contenedores de Docker Compose y así tener acceso a la misma red de la BD. Para levantar Redash se debe ejecutar el único comando descrito en la tabla a continuación:

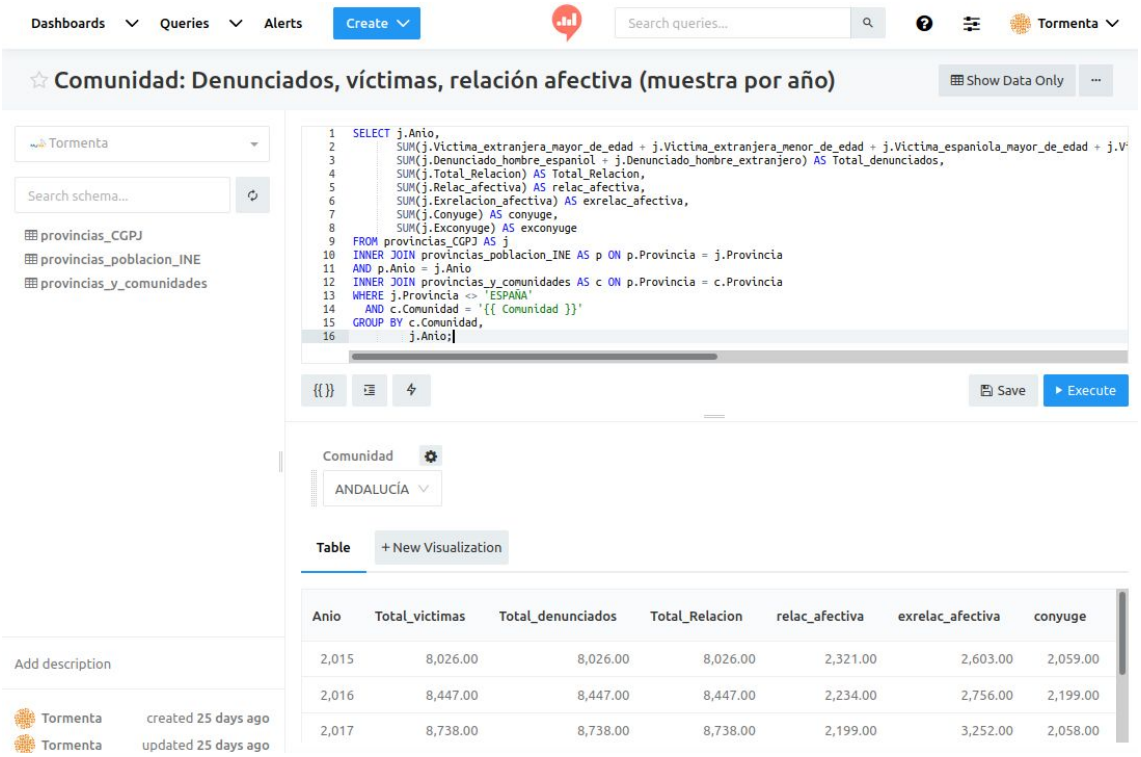
Tabla 4.4.2 Comando para iniciar Redash

Proceso	Comando	Observaciones
Iniciar **Redash**	`docker-compose -f docker-compose-redash.yml up`	Este comando levanta, en el entorno local, la imagen de Redash.

Fuente: elaboración propia

En el ordenador local, Redash está disponible en el puerto 5000. Tras realizar la configuración de la BD de Tormenta se pueden configurar las primeras consultas. La siguiente figura muestra, a la izquierda, las tablas de Tormenta. Al lado, en la sección superior, el espacio para escribir las consultas en lenguaje SQL y abajo aparecerán los resultados de la consulta.

Figura 4.4.9 Interfaz de Redash con la BD de Tormenta



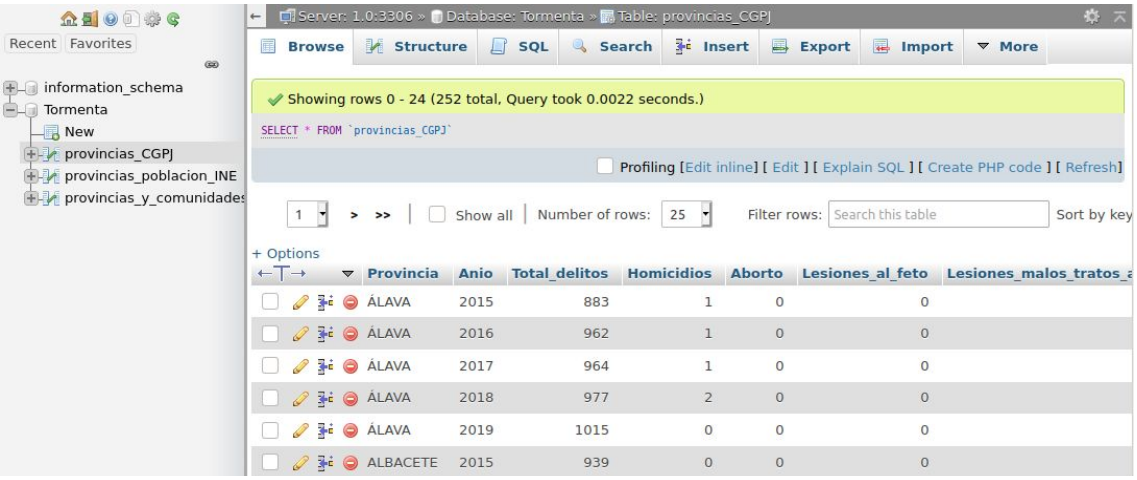
Fuente: Redash

5.Implantación

La implantación es la descripción de cómo se pone disponible la BD de Tormenta dentro de la herramienta de consulta y visualización de datos. Se comenta las herramientas elegidas para poner este proyecto a disposición a través de una dirección web. La instancia de Redash desplegada en DigitalOcean se encuenta disponible en la siguiente dirección IP: http://167.99.246.176/

Ya que el resultado del proceso ETL es una BD y este proceso, normalmente, se ejecuta una única vez después de haber superado la fase de testeo, no es necesario desplegar este código porque se puede ejecutar en un ordenador de forma local. Por lo tanto, el despliegue se limita a levantar una copia de la BD y a ponerla en marcha a través de Redash. Para desplegar la BD se ha optado por el proveedor Dinahosting, que ofrece una instancia de MariaDB de manera gratuita.

Figura 5.1 Interfaz de PHP Admin con los datos de tormenta



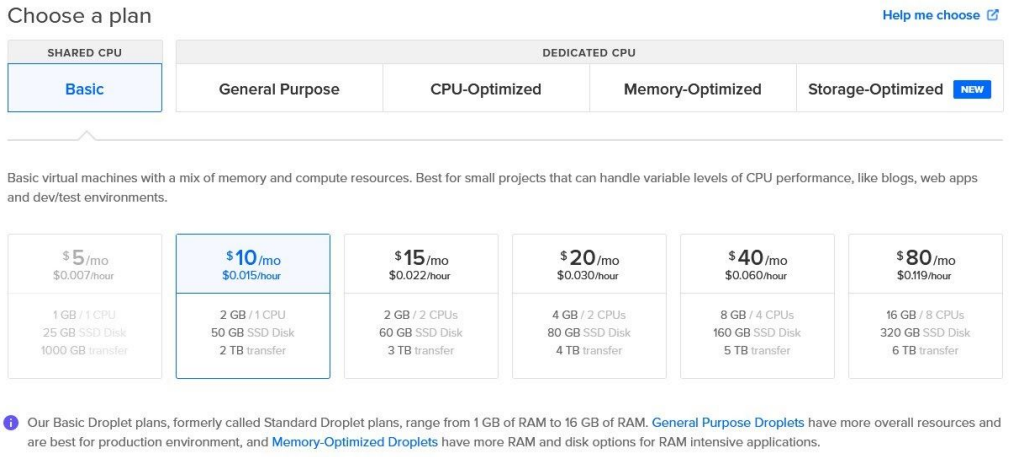
Fuente: Dinahosting

Como herramienta de visualización de los datos se opta por desplegar una instancia de Redash preconfigurada del *Marketplace* de DigitalOcean. Este proveedor se elige porque permite levantar una instancia de Redash con un único clic. La instancia se ha construído en el *Droplet*[20] más pequeño ofrecido, ya que por el uso que se le va a dar de momento es suficiente.

[20] Droplet es el nombre comercial que le da DigitalOcean a las máquinas virtuales que componen su *Cloud*.

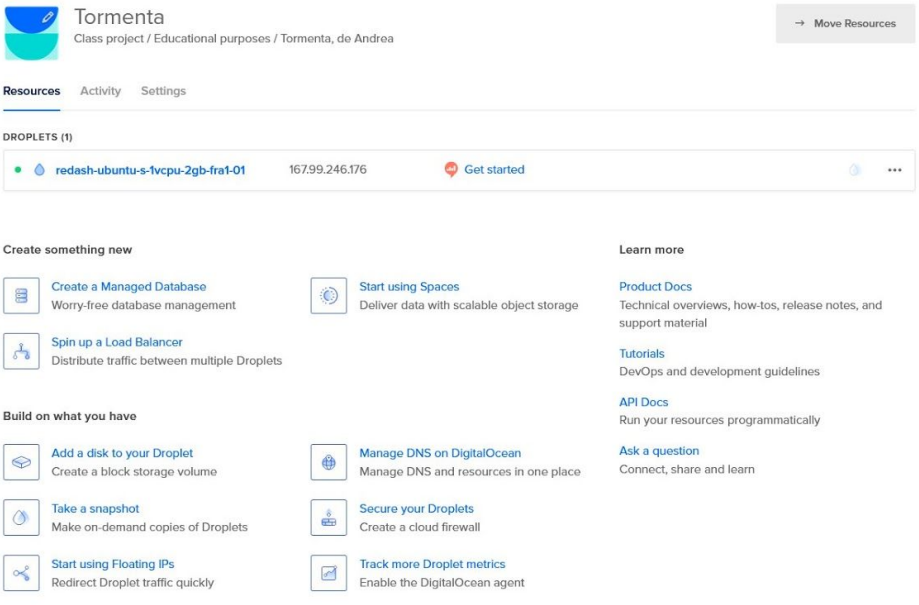
Tal como se puede ver en la siguiente figura, este droplet tiene un coste de €10 al mes. En caso de que su uso aumentase, DigitalOcean permite incrementar las capacidades del Droplet sin necesidad de reinstalar ni perder los datos, por ejemplo ampliar el disco duro disponible o hacer una copia de seguridad *(figura 5.3)*.

Figura 5.2 Diferentes tamaños de Droplets ofrecidos por DigitalOcean para Redash



Fuente: DigitalOcean

Figura 5.3 Operaciones que ofrece DigitalOcean en el Droplet de Redash



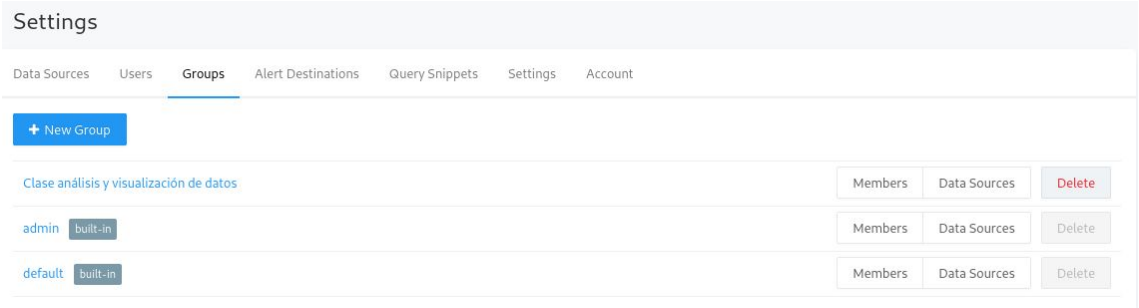
Fuente: DigitalOcean

6. Utilización de Tormenta

En este punto se prueba Tormenta y se realizan los primeros análisis estadísticos con el objetivo de mostrar los resultados y la rapidez con la que se puede acceder a los datos y realizar las consultas SQL. Para realizar los primeros análisis con Tormenta es necesario una introducción por la interfaz de Redash. En primer lugar se describirá la creación de grupos, usuarios y los accesos que se otorgan a través de ellos. Después, se explicará la interacción con Redash y la BD de Tormenta para usuarios con conocimientos del lenguaje de consulta SQL y para otros perfiles de usuarios.

En Redash se pueden crear usuarios y grupos. Un usuario puede pertenecer a distintos grupos y cada grupo puede tener diferentes permisos de acceso. La creación de grupos se puede observar en la siguiente figura, la interacción es similar para añadir un nuevo usuario. Solo los usuarios dentro del grupo llamado *admin* tienen permisos para editar y guardar la modificación de una consulta generada por otro usuario.

Figura 6.1 Interfaz para crear grupos y su listado



Fuente: adaptado de Redash con la BD de Tormenta

Redash tiene dos tipos de acceso a la BD: acceso limitado a las visualizaciones de los resultados de la consulta SQL junto con las gráficas, y acceso completo para copiar las consultas y generar nuevas gráficas. Solo el usuario llamado Tormenta puede modificar consultas de otros usuarios, porque es el único usuario administrador.

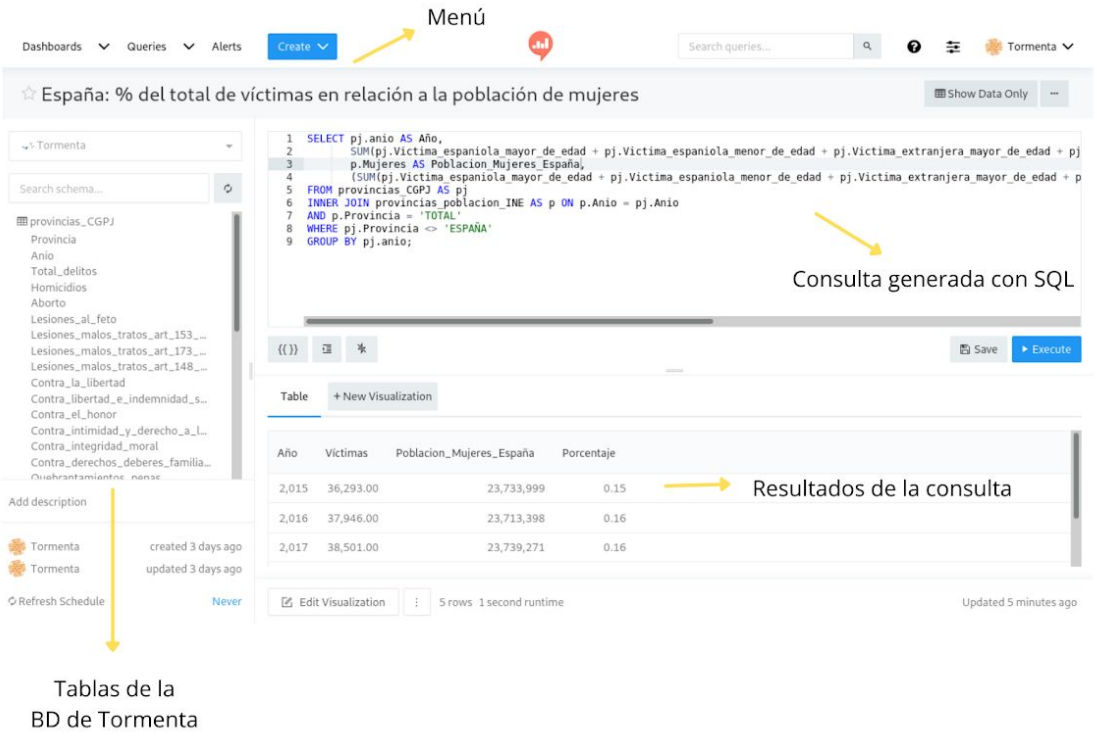
Los permisos mencionados anteriormente propician que un usuario pueda observar las consultas realizadas por otros usuarios. En la siguiente figura se puede observar una lista de consultas realizadas y los filtros que se puede hacer de ellas. En el caso de Tormenta, todos los usuarios tienen acceso de visualización de los resultados de la consulta, no necesariamente pueden observar la consulta SQL o crear una nueva.

Figura 6.2 Listado de consultas generadas por todos los usuarios

Fuente: adaptado de Redash con la BD de Tormenta

En la siguiente figura se muestra la interfaz para generar una consulta y visualizar los resultados o las gráficas. En la sección marcada como “Menú” se encuentra la lista de páginas para crear una nueva consulta o un nuevo escritorio. En “Tablas de la BD de Tormenta” se pueden visualizar las tablas y los nombres de las columnas, pero no los campos de las tablas. La consulta se debe crear en el apartado de “Consulta generada con SQL”. Los datos obtenidos se visualizan, después de ejecutar y guardar la consulta, en la sección inferior llamada “Resultados de consulta”.

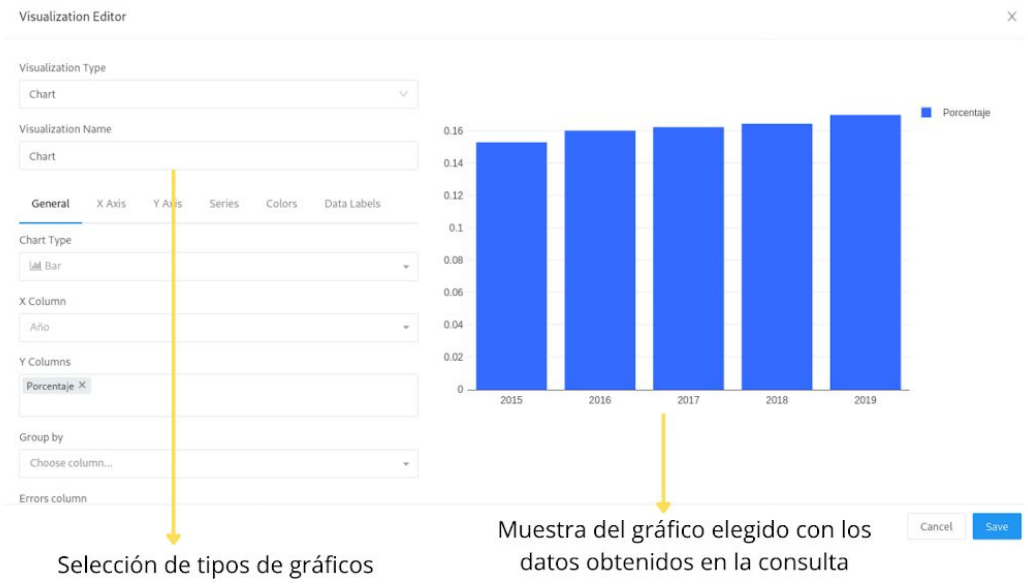
Figura 6.3 Descripción del escritorio de trabajo de Redash



Fuente: adaptado de Redash con la BD de Tormenta

Tras obtener los resultados de la consulta, estos se pueden visualizar con diferentes tipos de gráficos: gráficos de barras, de líneas, circulares, de áreas, de burbujas, entre otros. Algunos requieren que se indique lo que se desea visualizar. En la figura 6.5 se pueden observar algunos de los campos configurables. Los gráficos se visualizan mientras se elige la configuración. Al guardarlo permite otras interacciones como resaltar una parte del gráfico, agrandar una sección o descargarlo en formato PNG.

Figura 6.4 Interfaz para crear visualizaciones



Fuente: adaptado de Redash con la BD de Tormenta

Ahora bien, la BD de Tormenta con Redash puede ser utilizada por distintos perfiles que no tienen porqué conocer el lenguaje SQL. Para este tipo de usuarios se han creado 27 consultas iniciales, y los usuarios pueden realizar una copia de la consulta generando una nueva o con la opción *Fork* que creará una copia completa que incluye las visualizaciones que tenga creadas.

En la siguiente tabla se encuentra un índice de las 27 consultas, para que sean más fáciles de encontrar. Por ejemplo, la consulta con la numeración 1 contiene información del porcentaje de víctimas de una provincia en relación a la población de mujeres de la misma provincia, además la relación de estas con la cantidad de órdenes de protección iniciadas, tomando en cuenta que el filtro que se puede hacer es por cada año entre 2015 y 2019. En la figura 6.5 se puede ver este ejemplo.

Tabla 6.1 Índice de las 27 consultas iniciales de Tormenta

Descripción	Año		Pro.	Com.	País
	Prov.	Com.			
Porcentaje de denuncias en relación a la población de mujeres y víctimas con órdenes de protección en relación al total de denuncias	1	2	3	4	5

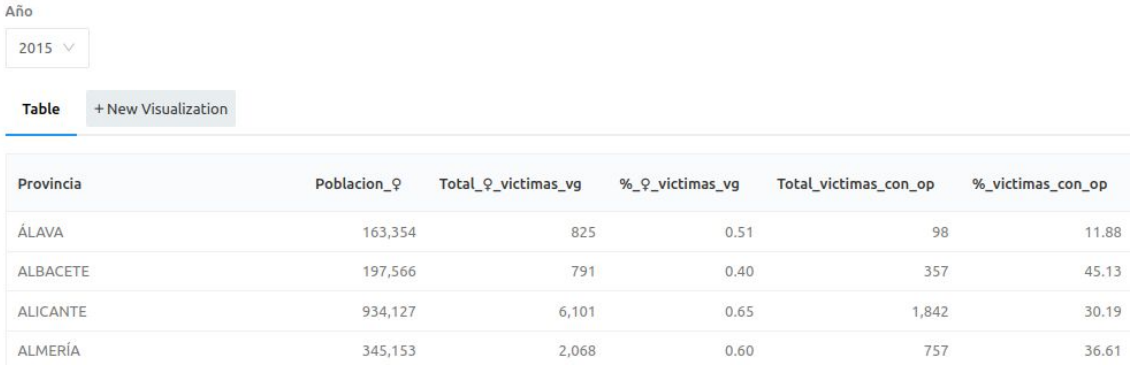
(continúa)

(continúa)

Descripción	Año		Pro.	Com.	País
	Prov.	Com.			
Porcentaje de denunciados, con órdenes de protección, en relación a la población de hombres	6	7	8	9	10
Porcentaje de víctimas con órdenes de protección: españolas, extranjeras mayores y menores de edad en relación a la población de mujeres	11	12	13	14	15
Porcentaje de denunciados con órdenes de protección: españoles y extranjeros en relación a la población de hombres	16	17	18	19	20
Porcentaje de acuerdo al tipo de relación (estos casos tienen órdenes de protección): cónyuge, ex cónyuge, relación afectiva y ex relación afectiva	21	22	23	24	25

Fuente: elaboración propia

Figura 6.5 Resultados de la consulta de provincias por porcentaje de víctimas en relación a la población de mujeres del año 2015.



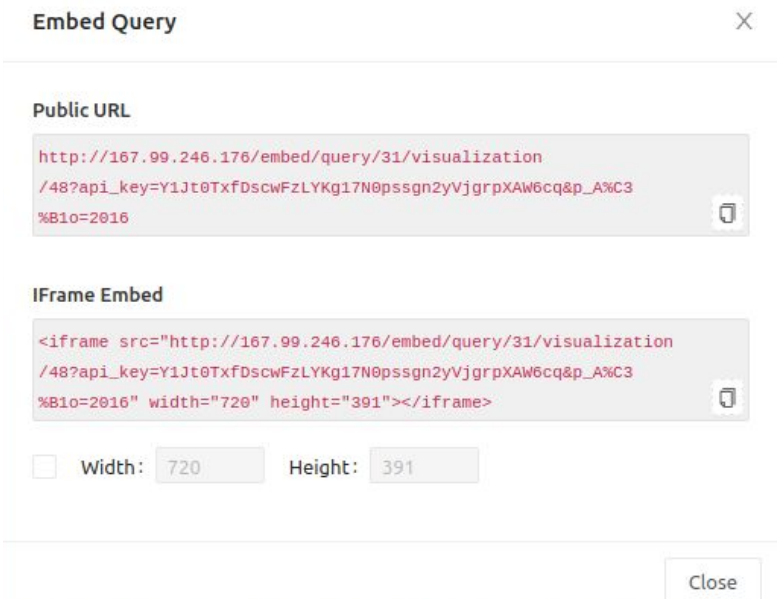
Provincia	Poblacion_♀	Total_♀_victimas_vg	%_♀_victimas_vg	Total_victimas_con_op	%_victimas_con_op
ÁLAVA	163,354	825	0.51	98	11.88
ALBACETE	197,566	791	0.40	357	45.13
ALICANTE	934,127	6,101	0.65	1,842	30.19
ALMERÍA	345,153	2,068	0.60	757	36.61

Fuente: adaptado de Redash con la BD de Tormenta

Otro uso que se le puede dar a Redash y Tormenta es la publicación de la información de las consultas en los formatos que Redash permite y que se pueden observar en la siguiente figura, estos son: CSV, archivo de Excel, con las opciones de Embed Elsewhere. Embed Elsewhere tiene 2 opciones, como se ve en la siguiente figura permite compartir la URL para que se pueda acceder a los resultados o gráficas. Un ejemplo de este uso está en la figura 6.6. La segunda opción es una etiqueta que se puede insertar dentro de una página web, al igual que un vídeo de Youtube. El ejemplo

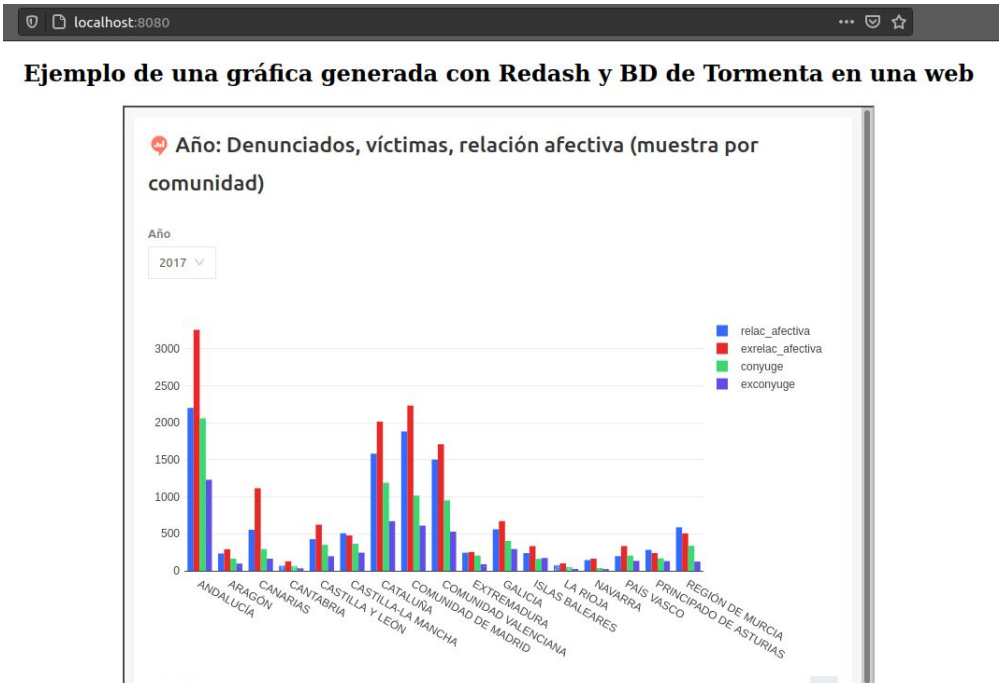
de esta opción está en la figura 6.7, en la que se puede observar una sencilla página web con la gráfica insertada y la posibilidad de interactuar con el filtro.

Figura 6.6 Opciones para compartir resultados y gráficas desde Redash



Fuente: adaptado de Redash con la BD de Tormenta

Figura 6.7 Ejemplo de la inserción de la etiqueta de la gráfica en una página web



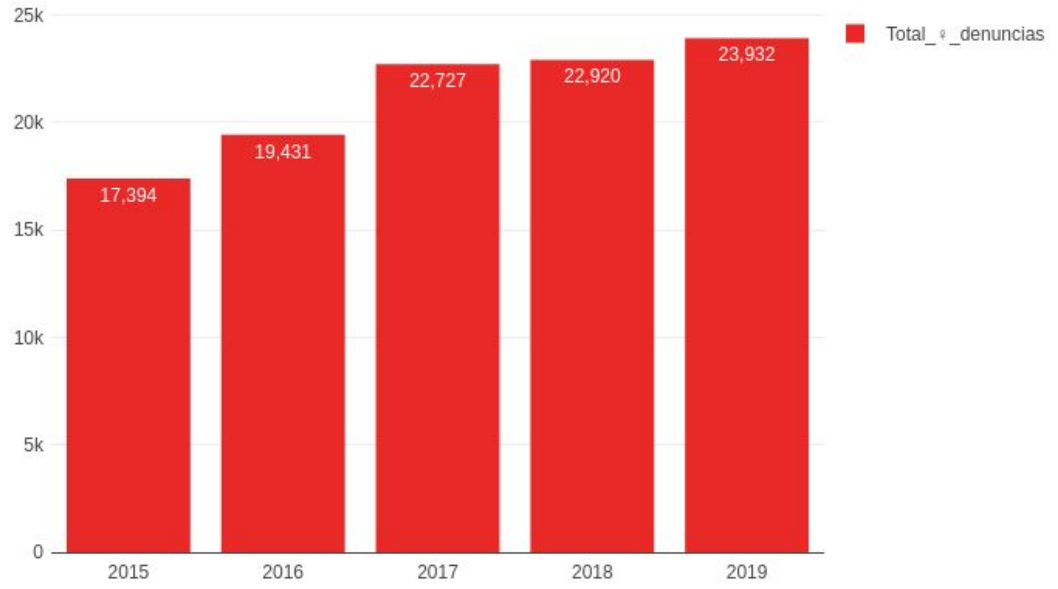
Fuente: adaptado de Redash con la BD de Tormenta

6.1. Análisis

A partir de las consultas anteriores se realiza el análisis de violencia de género desde los organismos judiciales en la Comunidad Valenciana, y se la compara en relación a España. En las gráficas se ha utilizado el acrónimo “op” para referirse a órdenes de protección. Es necesario aclarar que en los datos del CGPJ hace falta información para entender lo que se recoge en cada columna de los datos de sus archivos, por lo que se accedió a informes[21] anuales del mismo portal y al informe de Feminicidio.nat[22] de 2015, para obtener más información. A continuación se muestran las visualizaciones y resultados en el período comprendido entre 2015 y 2019.

En la Comunidad Valenciana se puede observar una tendencia al incremento de denuncias por violencia de género. Las denuncias pueden ser presentadas directamente por la víctima y por familiares en el juzgado o con un atestado policial, también están las denuncias que pueden ser por intervención policial directa. Como se puede apreciar en la siguiente gráfica, en el año 2019 todas estas denuncias llegaron a 23.932, lo que, para tener una idea más precisa del alcance, equivale a que más de la mitad de la población de mujeres de la provincia de Soria en 2019.

Gráfica 6.1.1 Evolución de denuncias por violencia de género en la Comunidad Valenciana



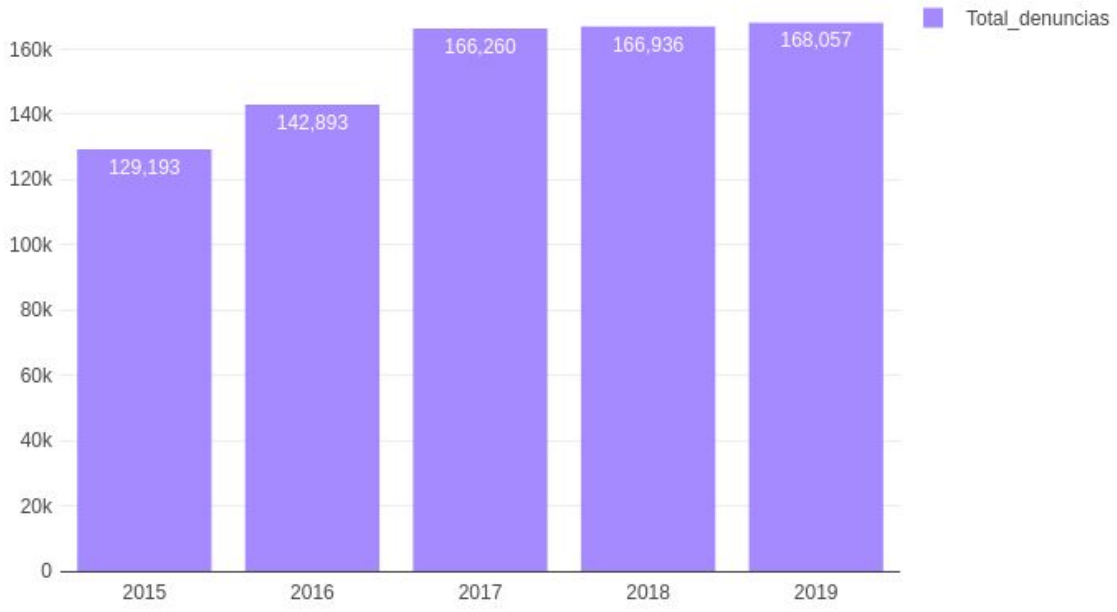
Fuente: adaptado de Redash con la BD de Tormenta

[21] El informe de 2015 del CGPJ de la violencia sobre la mujer se puede encontrar en: https://cutt.ly/XzTYfkH

[22] Feminicidio.net pertenece al dominio de internet feminicidi.cat y es parte de la asociación La Sur de Cataluña. Informe disponible en: https://cutt.ly/BzYdAi0

A nivel de España también se muestra un incremento de denuncias. En 2019, la cantidad de mujeres que se encontraron en un proceso legal fue equivalente a un poco más del total de la población de mujeres de La Rioja en 2019.

Gráfica 6.1.2 Evolución de denuncias de violencia de género de España



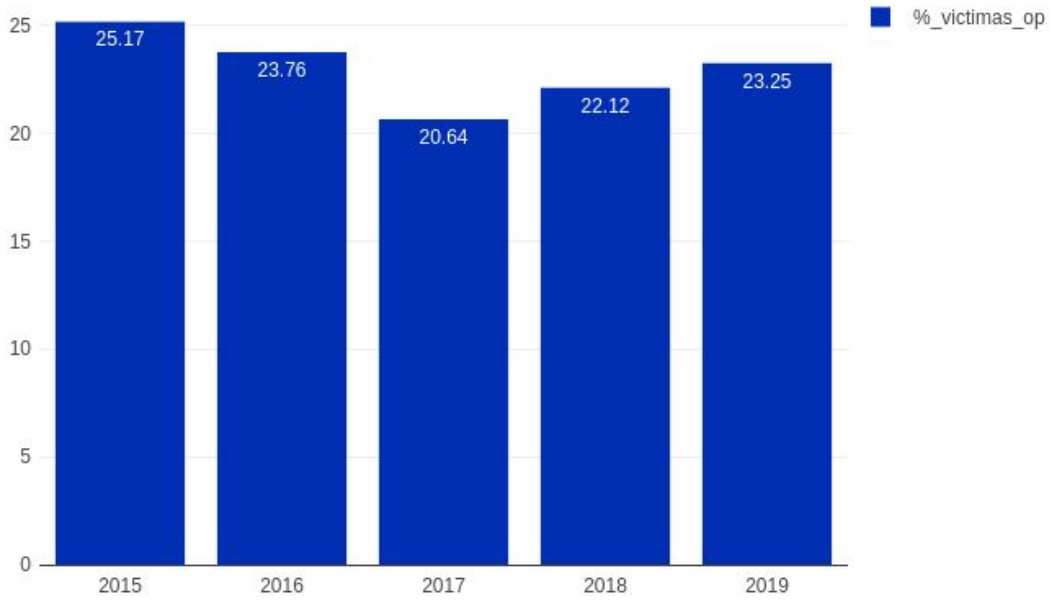
Fuente: adaptado de Redash con la BD de Tormenta

En las siguientes gráficas se puede apreciar que del total de denuncias, solo alrededor del 25% solicitan una orden de protección. Hay que tener en cuenta que los datos de órdenes de protección están relacionados con la cantidad de solicitudes que se inician, es decir, incluyen solicitudes de órdenes acordadas, denegadas o no admitidas. Esto quiere decir que en 2015 solo 1 de cada 4 mujeres, que tenía una denuncia, inició un proceso de solicitud de orden de protección. El resto de años, el porcentaje fluctúa entre un 23% y un 20%, por lo que en ocasiones ha llegado a ser 1 de cada 5 mujeres.

Gráfica 6.1.3 Denuncias y órdenes de protección de la Comunidad Valenciana

Fuente: adaptado de Redash con la BD de Tormenta

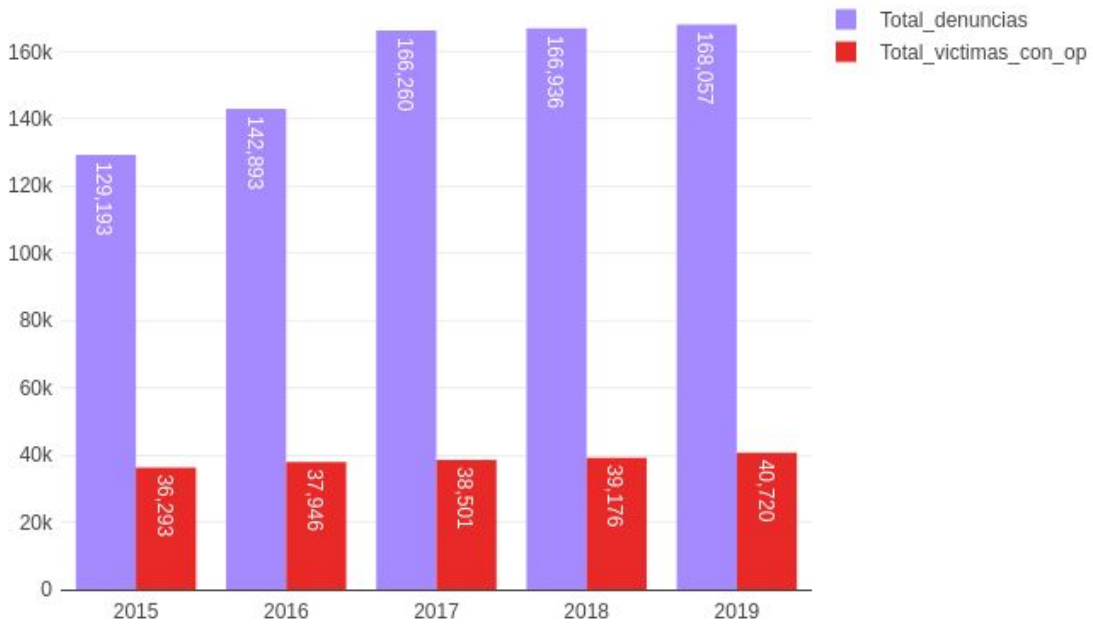
Gráfica 6.1.4 Porcentaje de órdenes de protección en relación al total de denuncias de la Comunidad Valenciana



Fuente: adaptado de Redash con la BD de Tormenta

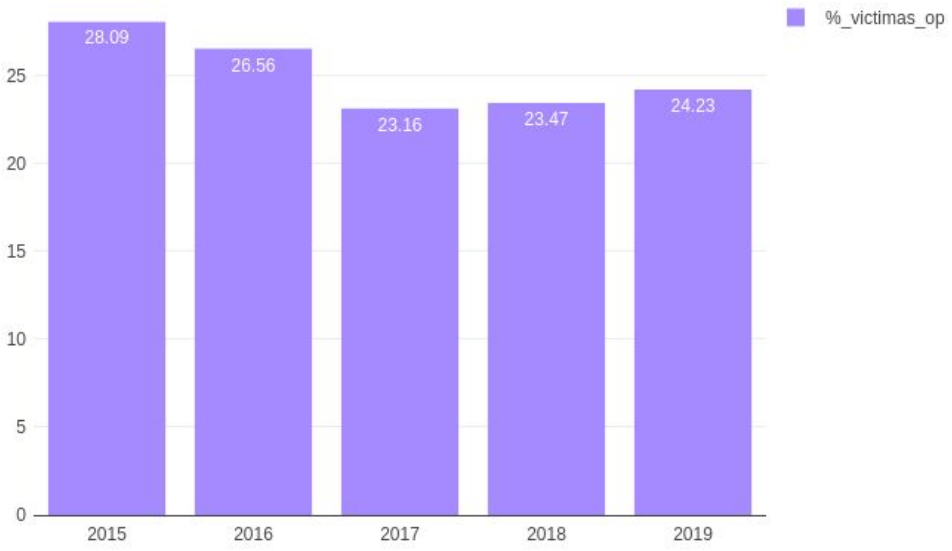
A nivel nacional, el porcentaje de solicitudes de órdenes de protección en relación a la cantidad de denuncias es similar al de la Comunidad Valenciana. Si bien las denuncias tienen una tendencia de aumento, las solicitudes de órdenes de protección fluctúan entre un 23% y un 28%, lo que nos indica que es alrededor de 1 de cada 4 mujeres con denuncia quienes han solicitado una orden de protección.

Figura 6.1.5 Denuncias y órdenes de protección de España



Fuente: adaptado de Redash con la BD de Tormenta

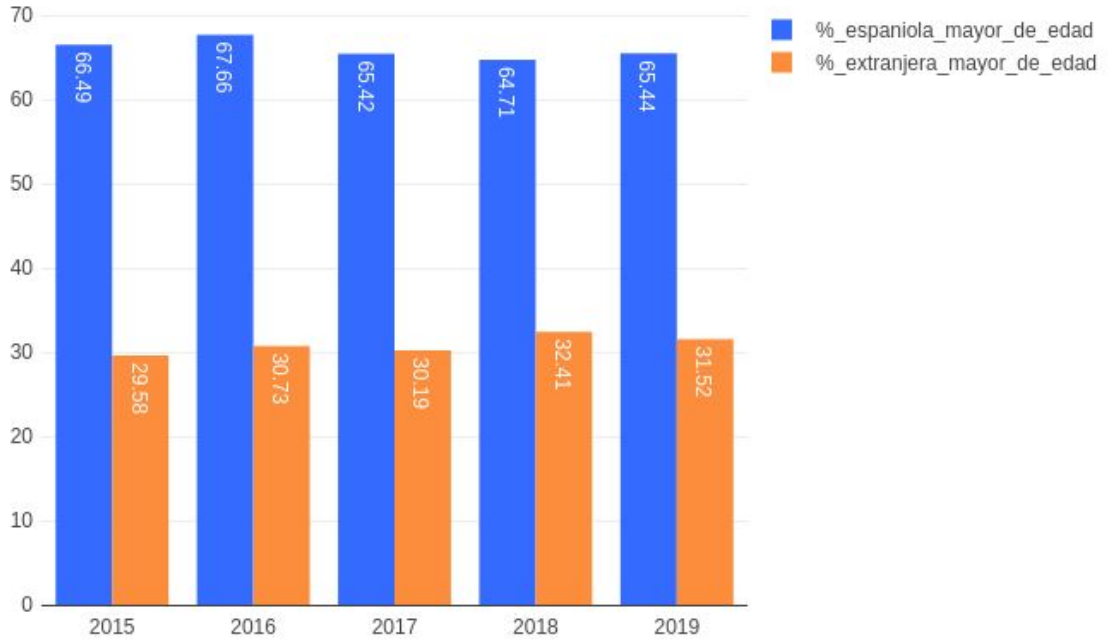
Gráfica 6.1.6 Porcentaje de órdenes de protección en relación al total de denuncias de España



Fuente: adaptado de Redash con la BD de Tormenta

Ahora bien, el inicio de un proceso de orden de protección, sin tomar en cuenta si se ha concedido, rechazado o ha existido una renuncia al proceso, es sobre todo realizado por mujeres españolas mayores de edad. De 2015 a 2019 se encuentran sobre el 65%. De esta forma, se puede inferir que 1 de cada 3 mujeres que han iniciado la solicitud de una orden de protección, es una mujer extranjera.

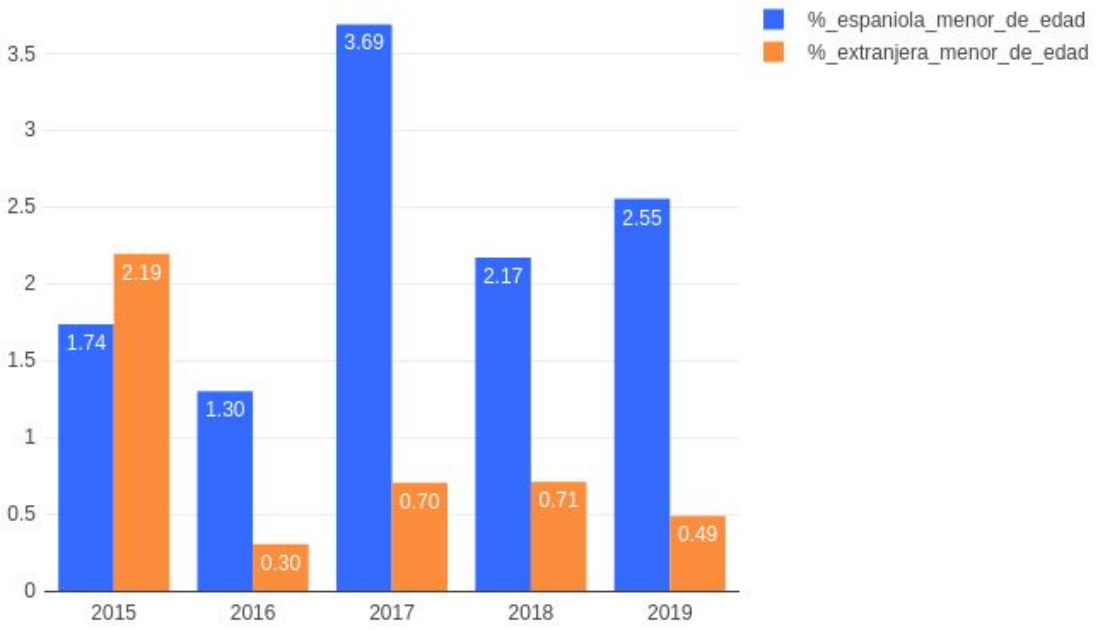
Gráfica 6.2.7 Porcentaje víctimas españolas y extranjeras mayores de edad en la Comunidad Valenciana



Fuente: adaptado de Redash con la BD de Tormenta

En el mismo sentido que la gráfica anterior, la relación entre mujeres españolas o extranjeras menores de edad que han iniciado una solicitud de orden de protección es muy bajo en relación a las mayores de edad. Por otro lado, las mujeres españolas menores de edad inician más procesos que las extranjeras, excepto en 2015 que es el único año en el que el 2.19% de mujeres extranjeras menores de edad lo inició mientras que las mujeres españolas fueron un 1.74%.

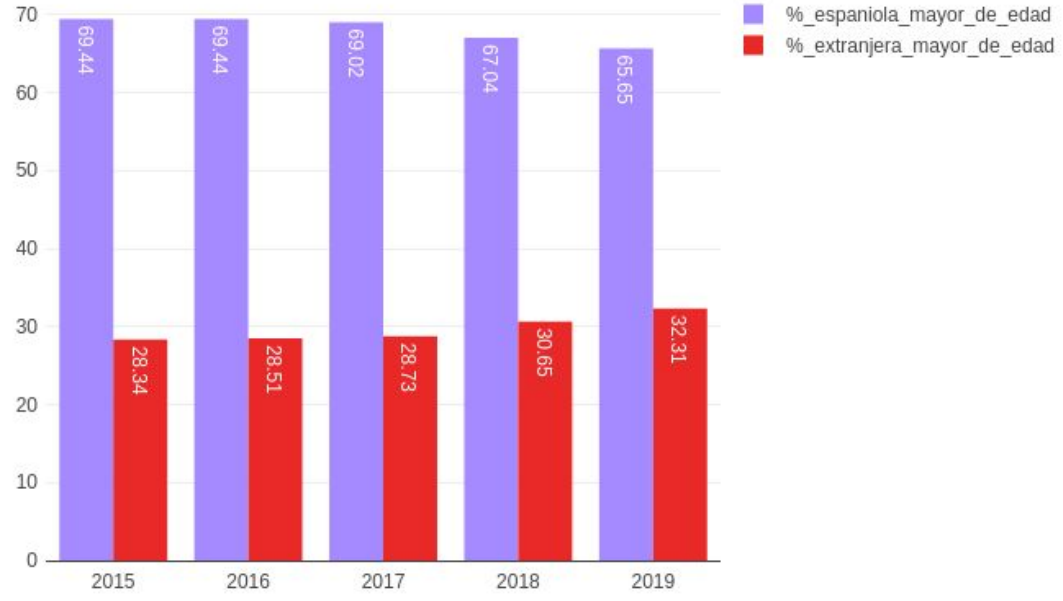
Gráfica 6.2.8 Porcentaje víctimas españolas y extranjeras menores de edad en la Comunidad Valenciana



Fuente: adaptado de Redash con la BD de Tormenta

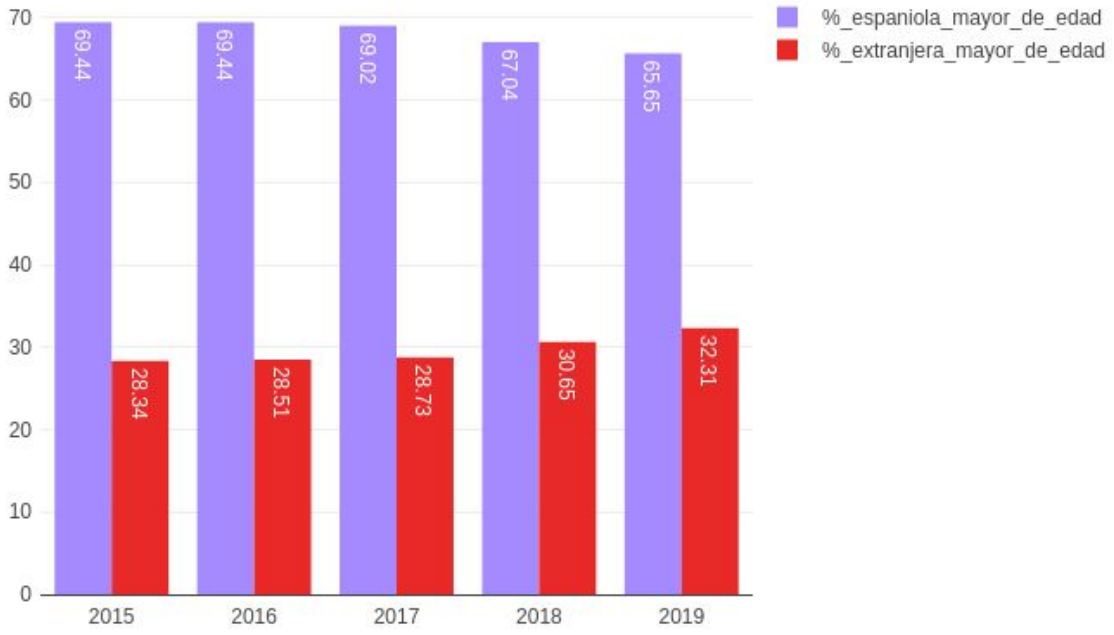
A nivel nacional, se mantiene que aproximadamente 2 de cada 3 mujeres que han iniciado una solicitud de orden de protección es una mujer española mayor de edad. Y en la gráfica 6.2.10 se puede observar que hay más solicitudes por parte de mujeres españolas menores de edad que de extranjeras menores de edad.

Gráfica 6.2.9 Porcentaje víctimas españolas y extranjeras mayores de edad en España



Fuente: adaptado de Redash con la BD de Tormenta

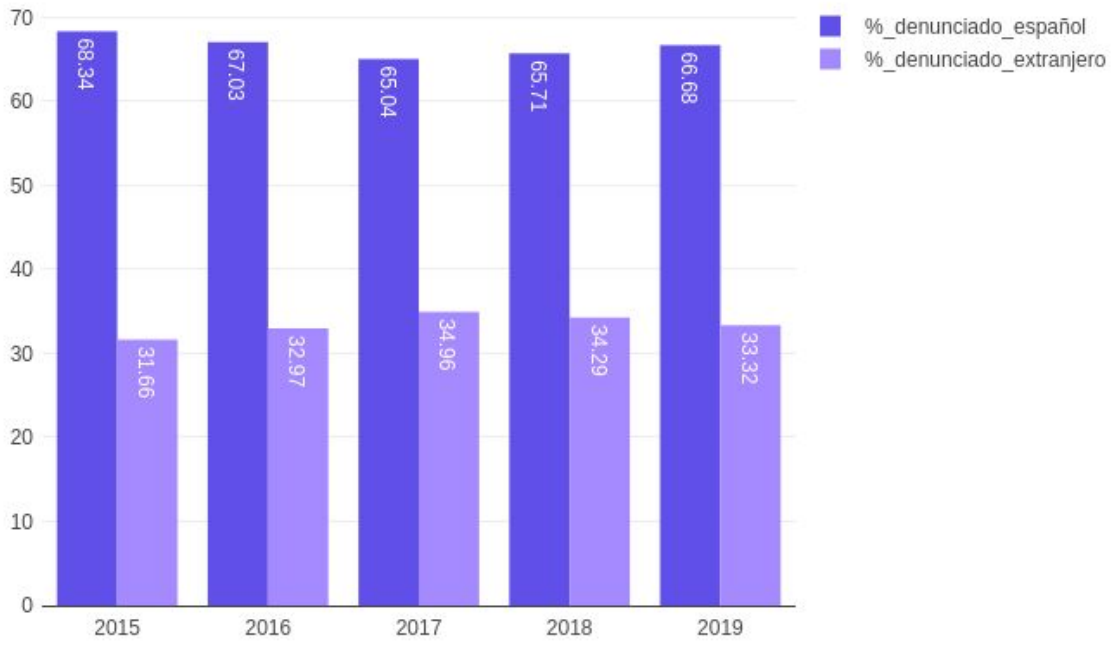
Gráfica 6.2.10 Porcentaje víctimas españolas y extranjeras menores de edad en España



Fuente: adaptado de Redash con la BD de Tormenta

Para los datos de hombres denunciados de acuerdo a su nacionalidad, también se debe tener en cuenta que son porcentajes relacionados con el total de casos con solicitudes de órdenes de protección, como antes, sin tener diferenciación por si han sido concedidas, denegadas o han tenido renuncias al proceso. Entonces, el total de hombres españoles denunciados entre 2015 a 2019 es superior al 65% en todos los años, estando alrededor de 2 de cada 3 denunciados, mientras que 1 es extranjero.

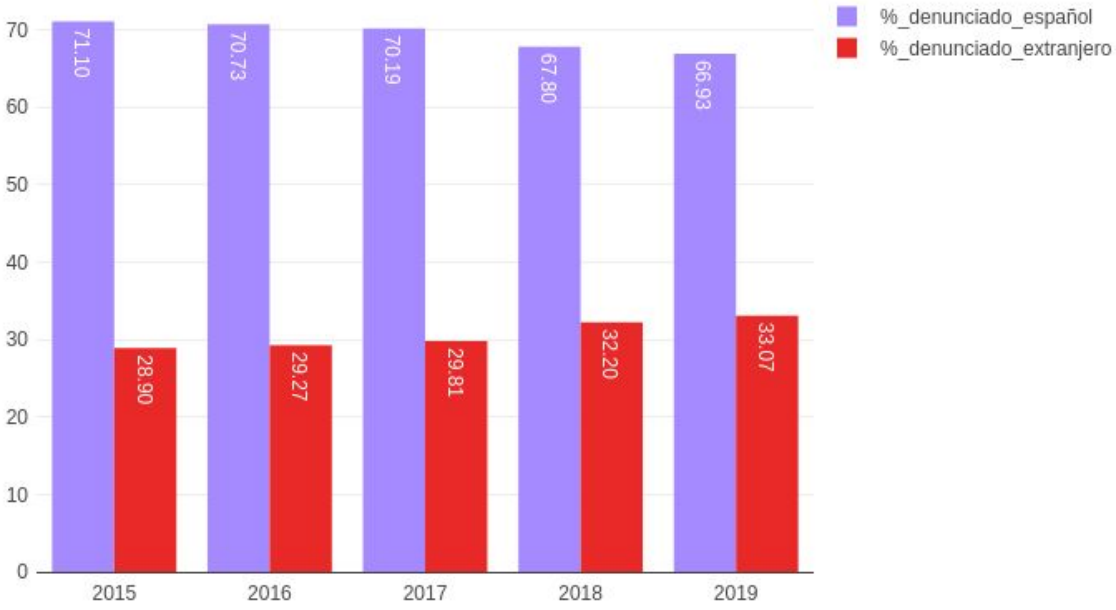
Gráfica 6.2.11 Porcentaje de denunciados españoles y extranjeros en la Comunidad Valenciana



Fuente: adaptado de Redash con la BD de Tormenta

A nivel nacional, en todos los años, los hombres españoles denunciados superan el 65%, por lo que los extranjeros no llegan al 35%. Estos porcentajes varían muy poco en la totalidad de los años estudiados.

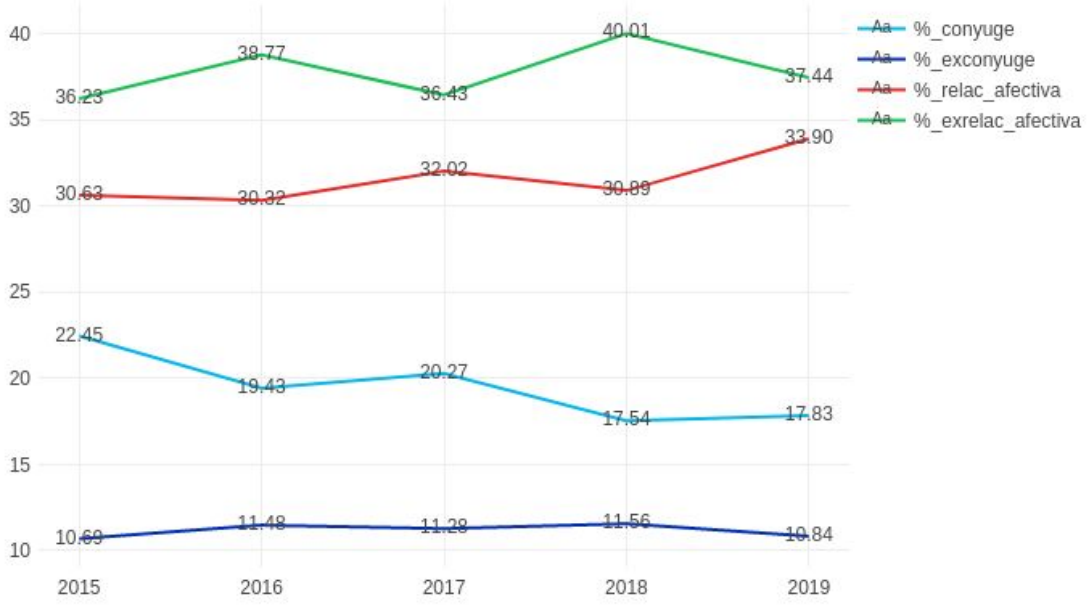
Gráfica 6.2.12 Porcentaje de denunciados españoles y extranjeros en España



Fuente: adaptado de Redash con la BD de Tormenta

Ahora bien, la mayor cantidad de denuncias con solicitudes de órdenes de protección, de acuerdo a la relación de parentesco, como se denomina en los informes anuales de violencia de género del CGPJ, se registran en ex relaciones afectivas, con unos porcentajes alrededor del 35% entre los años 2015 y 2019. Después se encuentran los casos en relaciones afectivas, que en todos los años sobrepasaron el 30%. En tercer lugar, las denuncias a cónyuges que están sobre el 17% y las denuncias a ex cónyuges sobre el 10%. De esta forma, los porcentajes más bajos se encuentran entre cónyuges o excónyuges, siendo los más altos en las ex relaciones afectivas, seguidos de las relaciones afectivas. Y haciendo un balance global, 2 de cada 3 casos se dieron por parte de hombres con los que las mujeres mantuvieron o mantienen una relación afectiva. Sin embargo, para ver si hay más maltrato en los matrimonios o en las relaciones afectivas, necesitaríamos saber la cantidad de estos tipos de relaciones en cada provincia, aunque intuimos que hay más parejas que matrimonios.

Gráfica 6.2.13 Porcentaje por tipo de relaciones: cónyuge y excónyuge en la Comunidad Valenciana



Fuente: adaptado de Redash con la BD de Tormenta

En España también se encuentran más denuncias con órdenes de protección solicitadas entre ex relaciones afectivas. Después se encuentran las solicitudes en casos por relaciones afectivas que superan el 30% en todos los años. Por su parte, las solicitudes con relaciones de cónyuges son más altas que entre ex cónyuges, aunque muestran una tendencia a la disminución, siendo 2015 el año en el que fueron el 24.38% de las denuncias para llegar al 2019 a ser el 18.72%. De nuevo, globalmente

es mayor el maltrato en los casos en las parejas y exparejas que en las maridos y ex-maridos.

Gráfica 6.2.15 Porcentaje por tipo de relaciones: relación afectiva y ex relación afectiva en España

Fuente: adaptado de Redash con la BD de Tormenta

7.Conclusiones

Tormenta utiliza los datos de violencia de género para mostrar el impacto de la centralización de datos y las relaciones establecidas entre ellos, sobre todo, en la disminución del tiempo al realizar análisis estadísticos por lo que se han alcanzado los objetivos iniciales de este TFM al cumplir con:

- Tener datos normalizados y centralizados de 2 fuentes de datos abiertos y una fuente de datos creada por ser de conocimiento público.
- Se han establecido relaciones entre las diferentes fuentes.
- Las principales relaciones son por el nombre de las provincias, comunidades y el año. Existen otras relaciones que se establecen en las consultas SQL, como el número de víctimas y denunciados.
- La BD y las consultas se han puesto a disposición en la dirección web http://167.99.246.176/, el único requisito, para acceder a ellas, es tener un usuario y contraseña.
- La utilidad de la centralización de datos ha tenido el mayor impacto en la reducción de tiempo con la que se puede realizar una análisis, porque depende del tiempo que se tarde en crear una consulta SQL. En el caso de utilizar una consulta creada, solo depende del filtro de búsqueda y de la visualización que se le quiera dar.
- La BD de Tormenta cuenta con una tabla de datos por cada fuente.
- La BD de Tormenta puede ser utilizada en cualquier otra herramienta de consulta y visualización.
- Para modificar o clonar el proyecto se necesita conocimiento informático de GitLab, Docker, Docker-Compose, JavaScript para el proceso ETL, y para realizar la conexión con Redash.
- El único servicio por el que paga Tormenta para poder ser utilizada por una dirección web es su servidor en Digital Ocean.

Este proyecto ha sido una muestra de la utilidad que puede tener la normalización y centralización de datos para una consulta en la que se establezcan relaciones. Pero es el inicio del potencial que tiene, por lo que las posibilidades de extensión se detallan en el apartado de trabajo futuros.

7.1. Relación del trabajo desarrollado con los estudios cursados

Este proyecto es el resultado del conocimiento adquirido en el Máster en Gestión de la Información y la formación dentro de Devscola, una asociación para aprender a programar de acuerdo a las buenas prácticas. A continuación se detallan las competencias transversales y las asignaturas del máster que se han requerido y puesto en práctica para la elaboración del presente trabajo académico.

En primer lugar, la aplicación y el pensamiento práctico por la búsqueda de una solución útil para reducir el tiempo que se suele invertir en proyectos de análisis de datos abiertos sobre la violencia de género. La solución ha sido pensada para usuarios que ya son requeridos por la sociedad de la información en la que se vive, a partir de la Revolución Industrial 4.0 o Revolución Digital.

En segundo lugar, el análisis y solución de problemas junto al conocimiento de problemas contemporáneos para ayudar a resolver un problema que afecta a gran parte de la población mundial. También que aporta en los ODS, por lo tanto en una solución para alcanzar un mundo más equitativo.

En tercer lugar, el diseño y proyecto al incorporar fases para la creación de Tormenta a partir del proceso ETL, además de automatizar el proceso de extracción y limpieza de datos a partir de los *secuencia de comandos*. También se debe tener en cuenta que para demostrar la utilidad del proyecto, se termina con un análisis final que prueba el beneficio de su uso.

Sobre el aprendizaje dentro de las aulas que han brindado la posibilidad, han sido instrumento y han otorgado las herramientas fundamentales para el desarrollo son: Explotación de datos masivos, Sociedad de la información y Almacenamiento y recuperación de información.

Explotación de datos masivos por los conocimientos para abordar la recuperación, gestión y explotación de los datos. Es decir, procesos ETL que permiten generar conocimiento útil desde los datos. Sociedad de la información por la importancia de la

creación de proyectos con impacto social, el conocimiento de los aportes desde la sociedad para cumplir con la Agenda 2030. Finalmente, Almacenamiento y recuperación de información porque se ha trabajado con información estructurada, tablas que tienen definidos sus campos, además de las bases de datos basadas en el modelo relacional SQL.

7.2. Trabajos futuros

Este proyecto no pretende ser el fin sino el inicio de mejoras para el tratamiento de datos de violencia de género, también de otros ámbitos en los que pueden impactar el uso de los datos abiertos y pueden reutilizar esta solución. A continuación se enlistan las posibilidades que podría tener Tormenta, se detallan algunas funcionalidades adicionales que podrían incorporarse en el futuro:

- Ya que actualmente la consulta de datos es abierta, pero no la modificación y la incorporación de datos, se podrían preparar *endpoints* para que otras personas puedan incorporar nuevas fuentes de datos. Hay que tener en cuenta que la funcionalidad de añadir datos es externa a Redash, que en este caso es la herramienta de consulta.
- La herramienta de consulta utilizada podría permitir la creación o configuración de usuarios con diferentes accesos de uso: consulta, modificación e incorporación de nuevos usuarios y datos.
- Si bien la incorporación de fuentes de datos es interesante, sería un aporte importante para la BD de Tormenta, que los usuarios puedan subir CSV con datos que ya han pasado por un proceso de limpieza manual.
- Aceptar código para que los usuarios suban nuevas secuencias de comandos que puedan ser reutilizados, modificados y mejorados, que se utilicen en limpieza y tratamiento de datos como el que se está implementando en Tormenta.
- Automatizar la descarga y actualización de datos de manera periódica.
- Realizar un sistema de creación automática de tablas, porque actualmente las tablas se crean de forma manual.
- Desarrollar una interfaz de usuario para que crear los archivos de configuración no sean a través de un archivo sino de una interfaz gráfica, lo que puede acercar esta herramienta a personas que no tienen conocimiento de SQL.

8.Referencias

Ardévol, E. (2017) BIG DATA Y DESCRIPCIÓN DENSA, Virtualis, publicación editada por el Instituto. Tecnológico y de Estudios Superiores de Monterrey

Ávila, R. (2020, 11 6). *DecidimFest20: Renata Ávila* [video]. Youtube. Recuperado 11 6, 2020, de https://youtu.be/8dYUa1XdMwo?t=995

Barcelona Iniciativa Open Data. (n.d.). *Fases del proyecto Datos x Violencia x Mujeres*. Recuperado 11 19, 2020, de http://www.datosviolenciamujeres.es/fases-datos-violencia-genero-proyecto/

Blanco Castilla, E. P., & Quesada, M. (2016). *Periodismo de datos*. Javier Herrero y Milena Trenta. 10.4185/cac112

Bustamante Martínez, A., Galvis Lista, E. A., & Gómez Flores, L. C. (2013). Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI. *Scientia et Technica*, *18*(1), 185 - 191. https://doi.org/10.22517/23447214.8727

Celma Giménez, M., Casamayor Cárdenas, J. C., & Mota Herranz, L. (2002). *Bases de datos relacionales* ((18/01/2002) ed.). Universidad politécnica de Valencia.

CGPJ. (n.d.). *Violencia doméstica y de género, Actividad del Observatorio, Datos estadísticos*. Datos estadísticos. Recuperado 12 3, 2020, de http://www.poderjudicial.es/cgpj/es/Temas/Violencia-domestica-y-de-genero/Actividad-del-Observatorio/Datos-estadisticos/

CGPJ. (2015, 11). *Diez años de la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género en los órganos judiciales [Boletín de información Estadística No 43]*. Recuperado 01 23, 2021, de www.poderjudicial.es/stfls/ESTADISTICA%20JUDICIAL%20NUEVO/FICHEROS/Datos%20de%20Justicia/Boletines%20Anteriores/Bolet%C3%ADn%20n%C2%BA%2043%20-%20Ley%20Proteccion%20Integral.pdf

Digital Fems. (n.d.). *Visibilizando las violencias machistas*. Datos contra el ruido. Recuperado 11 25, 2020, de https://datoscontraelruido.org/

Escolar, C. (2002). *El Proceso de "Gestión de Datos". Construcción, medición y evaluación de los datos*. Cinta de Moebio. https://www.redalyc.org/articulo.oa?id=101/10101404

Gerardo, C. G. (2008). *Un Sistema para el Mantenimiento de Almacenes de Datos*. Universitat Politècnica de València. https://doi.org/10.4995/Thesis/10251/2505

Guevara, C. (n.d.). ¿Por qué capacitar en género? Acercamiento a la problemática y a conceptos que aborda el Programa ELVG. *Incluir la mirada de género en la escuela*. http://schole.isep-cba.edu.ar/incluir-la-mirada-de-genero-en-la-escuela/4/

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). *ACM Human Factors in Computing Systems (CHI)*. Wrangler: Interactive Visual Specification of Data Transformation Scripts. http://vis.stanford.edu/papers/wrangler

Ley 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género. *Boletín Oficial del Estado* (103). 28 de enero de 2005. ISSN 0212-033X. BOE-A-2004-21760.

Logicalis. (2015, 02 22). Data hub: los nuevos sistemas de gestión de datos. Recuperado 11 27, 2020, de https://blog.es.logicalis.com/analytics/data-hub-las-nuevos-sistemas-de-gestion-de-datos

Naciones Unidas. (n.d.). *Objetivos de desarrollo sostenible*. Recuperado 11 09, 2020, de https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Peiró, B. (2019). *Periodismo, Datos y Objetivos de Desarrollo Sostenible. Informe Diagnóstico del Periodismo de Datos Sobre Objetivos de Desarrollo Sostenible en el País Valencià.* Valencia, UGT País Valencià, Instituto Sindical de Cooperación al Desarrollo y Generalitat Valenciana. https://www.ugt-pv.es/2018/2019/02/periodismo_datos_cast.pdf

Pollock, R., & Kariv, A. (n.d.). *About DataHub*. We want to make data better, together. Recuperado 11 27, 2020, de https://datahub.io/docs/about

Ruvalcaba Gómez, E. (2020). Datos abiertos. *Revista en Cultura de la Legalidad*, *18*, 327-334. https://doi.org/10.20318/eunomia.2020.5280

Varela, N. (2014). In *Feminismo para principiantes* (Edición digital ePub base r1.2 ed., p. 305). Titivillus.

Wikipedia. (n.d.). *Comma-separated values*. Recuperado 11 25, 2020, de https://en.wikipedia.org/wiki/Comma-separated_values

Anexos

Normalización de comunidades y provincias

Listado de comunidades y provincias con las normalizaciones de datos. A la izquierda se encuentran los nombres utilizados y a la derecha las variaciones encontradas.

Comunidades

- Andalucía: ANDALUCÍA,
- Aragón: ARAGÓN,
- Canarias: CANARIAS,
- Cantabria: CANTABRIA,
- Castilla y león: CASTILLA Y LEÓN,
- Castilla-La Mancha: CASTILLA- LA MANCHA,
- Cataluña: CATALUÑA/CATALUNYA,
- Comunidad de Madrid: COMUNIDAD DE MADRID,
- Extremadura: EXTREMADURA,
- Galicia: GALICIA,
- Islas Baleares: ILLES BALEARS,
- La Rioja: LA RIOJA,
- Región de Murcia: REGIÓN DE MURCIA,
- Navarra: COMUNIDAD FLORAL DE NAVARRA,
- País Vasco: PAÍS VASCO/EUSKADI,
- Principado de Asturias: PRINCIPADO DE ASTURIAS,
- Comunidad Valenciana: COMUNITAT VALENCIANA,

Provincias

- La Coruña: A CORUÑA, CORUÑA, A,
- Alicante: ALACANT/ALICANTE, ALICANTE/ALACANT,
- Albacete: ALBACETE,
- Almería: ALMERÍA,
- Ávila: ÁVILA,
- Asturias: ASTURIAS,
- Álava: ARABA/ÁLAVA, ARABA,

- Badajoz: BADAJOZ,
- Barcelona: BARCELONA,
- Burgos: BURGOS,
- Cáceres: CÁCERES,
- Cádiz: CÁDIZ,
- Cantabria: CANTABRIA,
- Castellón: CASTELLÓ/CASTELLÓN, CASTELLÓN/CASTELLÓ,
- Ceuta: CEUTA,
- Ciudad Real: CIUDAD REAL,
- Córdoba: CÓRDOBA,
- Cuenca: CUENCA,
- Guipuzcoa: GIPUZKOA, GUIPUZKOA,
- Girona: GIRONA,
- Granada: GRANADA,
- Guadalajara: GUADALAJARA,
- Huelva: HUELVA,
- Huesca: HUESCA,
- La Rioja: LA RIOJA, RIOJA, LA,
- Las Palmas: LAS PALMAS, PALMAS, LAS,
- Lérida: LLEIDA,
- Lugo: LUGO,
- Madrid: MADRID,
- Melilla: MELILLA,
- Murcia: MURCIA,
- Navarra: NAVARRA,
- Ourense: OURENSE,
- Palencia: PALENCIA,
- Pontevedra: PONTEVEDRA,
- Salamanca: SALAMANCA,
- Santa Cruz de Tenerife: SANTA CRUZ DE TENERIFE,
- Segovia: SEGOVIA,
- Sevilla: SEVILLA,
- Soria: SORIA,
- Tarragona: TARRAGONA,
- Teruel: TERUEL,
- Toledo: TOLEDO,

- Valladolid: VALLADOLID,
- Zamora: ZAMORA,
- Zaragoza: ZARAGOZA,
- Islas Baleares: ILLES BALEARS, BALEARS, ILLES, ILLES BALEARES,
- Jaén: JAÉN,
- León: LEÓN,
- Málaga: MÁLAGA,
- Valencia: VALÈNCIA/VALENCIA, VALENCIA/VALÈNCIA,
- Vizcaya: BIZKAIA