Facultad de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRADO
GRADO EN INGENIERÍA INFORMÁTICA
MENCIÓN EN SISTEMAS DE INFORMACIÓN

Integración y Análisis de datos para evaluar restricciones COVID-19

Estudiante: Béaz Brenlla Miguel
Dirección: Susana Ladra González
Dirección: Fernando Silva Coira

A Coruña, novembro de 2021.

A mi madre, por todo.

Agradecimientos

Especial agradecimiento a mi madre, por estar ahí en todo momento, mi padre por la confianza y a mis abuelos. También agradecer a los diferentes profesores que he tenido hasta el día de hoy, sin los cuales no podría haber tenido los conocimientos necesarios para llegar a donde estoy. Y finalmente a mis amigos y personas más cercanas por los ánimos y el apoyo.

Resumen

Con este proyecto se busca realizar una investigación sobre la pandemia con el objetivo de saber cómo ha afectado esta al territorio gallego y ver cómo han funcionado las medidas tomadas para frenar los contagios y muertes propiciadas por la misma.

Para poder realizar ese análisis, es necesario seguir una serie de pasos que permitirán que el proyecto tenga una buena estructura y, por ende, poder elaborar informes de calidad.

En primer lugar se realizará una investigación enfocada en identificar qué se quiere analizar y, en consecuencia, qué datos harán falta para el análisis que se quiere desarrollar.

Tras este primer paso, se abordará la tarea de diseño y creación de un Data Warehouse con su respectivo proceso ETL. Para llevar a cabo la creación del almacén de datos se utilizará PostgreSQL (pgAdmin) y para crear los procesos ETL, se hará uso de Pentaho. Una vez construido el Data Warehouse, se crearán informes (mediante la herramienta Power BI) que permitirán visualizar los datos de la pandemia de una forma sencilla y detallada a la vez. Esto permitirá poder sacar conclusiones más claras sobre las consecuencias del COVID-19 en Galicia y el impacto que han tenido las restricciones.

Abstract

The aim of this project is to carry out a research on the pandemic in order to know how it has affected the Galician territory and to see how the measures taken to curb the contagions and deaths caused by it have worked.

In order to carry out this analysis, it is necessary to follow a series of steps that will allow the project to have a good structure and, therefore, to be able to produce quality reports.

First of all, research will be carried out to identify what is to be analyzed and, consequently, what data will be needed for the analysis to be carried out.

After this first step, the task of designing and creating a Data Warehouse with its respective ETL process will be addressed. PostgreSQL (pgAdmin) will be used to create the Data Warehouse and Pentaho will be used to create the ETL processes. Once the Data Warehouse is built, reports will be created (using the Power BI tool) that will allow us to visualize the pandemic data in a simple and detailed way at the same time. This will allow us to draw clearer conclusions about the consequences of the consequences of the COVID-19 in Galicia and the impact that the restrictions have had.

Palabras clave:

- COVID-19
- Power BI
- Almacén de datos
- ETL
- Pentaho

Keywords:

- COVID-19
- Power BI
- Data Warehouse
- ETL
- Pentaho

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1 Motivación

En estos últimos meses hemos vivido en nuestras carnes lo que es una pandemia y todo lo que conlleva. Nuestras vidas han dado un giro de 180º y nos encontramos en una situación que hace unos años era impensable. Pero, con la ayuda de los datos y la información que ahora se tiene sobre el COVID-19, podemos ver que hay una mejoría y se sabe cómo actuar ante casos que hace un año no se sabían abordar.

Los datos se han convertido en un factor muy importante y vital para prácticamente todos los ámbitos. Con ellos se pueden realizar estudios que permitan saber, de la forma más adecuada posible (más detallada, más resumida, esquemática, etc.), qué está pasando en la actualidad, en el pasado e incluso qué podría llegar a pasar en el futuro.

Todos los datos recopilados pueden tener múltiples funciones dependiendo de su enfoque. Por ejemplo, si se quiere saber el mejor tiempo en los 100 metros lisos de las últimas olimpiadas, simplemente se mostraría el dato del ganador de este año en la categoría. Pero si se quiere analizar si ese mismo competidor es considerado rápido dentro de los corredores de la historia, este dato habría que compararlo con los de los demás ganadores de la historia de las olimpiadas. De esta forma, el mismo dato, muestra dos posturas diferentes. Es muy importante utilizar los datos de la forma correcta para que estos sean útiles y se puedan realizar análisis y tomar mejores decisiones gracias a ellos.

Un claro ejemplo de su importancia es que, a la hora de enfrentar problemas/retos nuevos (de cualquier tipo), éstos suponen una mayor dificultad debido a que no se posee información previa de situaciones similares que ayude a abordarlos.

Algo similar ha pasado con el COVID-19. Durante los primeros meses, al no tener información previa de la enfermedad, todo era un caos y no se sabía cómo actuar ante el enorme problema. En la actualidad, con los datos recopilados en estos casi dos años, se sabe cómo actuar mejor ante diversas situaciones propiciadas por el virus.

Muchos pacientes se han salvado gracias a que otros (con sus mismos síntomas, gravedad, etc.) han permitido que se recopile, información útil y de calidad (qué tratamiento funciona mejor o peor, el tiempo del tratamiento, etc.) permitiendo que se actúe de forma más eficiente.

Existen varios factores importantes en la lucha contra el COVID-19, entre ellos se encuentran las restricciones (cuáles aplicar, cuándo hacerlo, cuándo quitarlas, para qué regiones...).

En este estudio, con respecto a las restricciones, se analizará la eficacia de las mismas y sus consecuencias.

Además del estudio del uso e impacto de las restricciones, también se analizará cómo ha afectado el COVID-19 a la población gallega: qué grupos de personas son los más afectados, cuáles tienen mayor o menor posibilidad de fallecer a causa de la enfermedad, si el sexo de la persona puede ser un factor importante a la hora de determinar la probabilidad de fallecer, etc.

Todo esto lo veremos en el estudio que se realizará a lo largo de este proyecto.

Para poder realizar ese análisis, es necesario seguir una serie de pasos que permitirán que el proyecto tenga una buena estructura y poder realizar informes de calidad.

En primer lugar se realizará una investigación enfocada a identificar qué se quiere analizar ya que la pandemia ha dejado una gran cantidad y variedad de datos que pueden ser procesados para llevar a cabo distintos análisis.

Una vez aclarado este punto, se procederá al diseño y creación de un Data Warehouse con su respectivo proceso ETL. Este último, permitirá utilizar datos provenientes de diferentes fuentes y realizar modificaciones y cargas de forma automatizada, mucho más sencilla que de forma manual.

La idea que se persigue con el proceso ETL automatizado es que los datos provengan de archivos CSV, que son muy sencillos de modificar (sin conocimientos informáticos avanzados), y poder realizar la actualización de los datos del Data Warehouse con un simple clic en una interfaz gráfica.

Una vez construido y cargado el Data Warehouse con los datos adecuados, se realiza la parte de reporting. En esta etapa se elaboran informes de diferentes tipos usando la información del almacén de datos.

Para cada informe habrá que estudiar bien la información que es relevante, indicar el por qué del mismo y qué pretende abordar. El fin que tienen es permitir que se saquen conclusiones sobre cómo afectó el COVID-19 a la población gallega y qué impacto tuvieron algunas de las restricciones que se han ido implantando a lo largo de la pandemia.

1.2 Objetivos

El objetivo que persigue este proyecto es conocer el impacto de la pandemia en Galicia y analizar los resultados de la implantación de algunas restricciones en las diferentes provincias o áreas sanitarias de la comunidad autónoma a través de informes.

De esta forma se podrá entender el poder que tienen los datos hoy en día y cómo es el proceso que estos deben seguir para poder ser utilizados de una forma adecuada permitiendo realizar estudios fiables.

Para poder llegar al paso final del proyecto (el propio análisis de los datos) se tienen que realizar una serie de pasos previos. Con ellos se abordarán diversos temas de interés que permitirán ver en profundidad los siguientes puntos:

- Diseñar y crear un Data Warehouse desde cero teniendo en cuenta qué se quiere almacenar en él y cómo van a ser usados sus datos.
- Creación de proceso ETL que adecue los datos de diferentes orígenes para poder recogerlos en el Data Warehouse de forma que sea posible realizar el estudio que se desea.
- Alcanzar cierta automatización del proceso ETL para facilitar la tarea de actualización de datos.
- Creación de informes que permitan analizar los datos de una forma más sencilla y visual con el fin de obtener conclusiones, tomar decisiones, etc.

1.3 Estructura de la memoria

A lo largo de este documento se explica cómo se ha llevado a cabo el Trabajo de Fin de Grado de forma detallada.

Se comienza con una breve explicación del objetivo que persigue el proyecto, la planificación del mismo y, acto seguido, se trata el apartado de definiciones de conceptos clave que serán necesarios para comprender todos los pasos del trabajo (Capítulo 2). Posteriormente se describen las metodologías utilizadas (Capítulo 3), desembocando en la explicación de cómo se ha desarrollado el proyecto (paso a paso) desde las fases iniciales en las que se diseña y se crea el Data Warehouse y los procesos ETL (Capítulo 4,Capítulo 5) hasta el análisis de los resultados ofrecidos por los informes (Capítulo 6). Finalmente se expresarán las conclusiones a las que el estudio ha dado pié además de unas posibles ideas para una futura continuación del proyecto (Capítulo 7).

Capítulo 2

Definiciones

A continuación, se verán algunos conceptos interesantes para poder comprender el proyecto, tanto sus pasos como su fin.

2.1 Data Warehouse

A pesar de su importancia, no hay que prestar atención únicamente a los datos, los lugares en los que se almacenan son igual de importantes. Es posible que esto sea obvio pero, en ocasiones, parece que la calidad del diseño del Data Warehouse y su estructura quedan en un segundo plano cuando es algo que se debe evitar a toda costa.

Pero... ¿qué es un Data Warehouse o Almacén de Datos?

Básicamente hace referencia a un sistema que combina información que proviene de diferentes fuentes, aunándola para que pueda ser usada con fines, normalmente, analíticos. Otra forma de definirlo sería una colección de datos [1] (no volátil, integrado y que puede variar en el tiempo) que proporciona apoyo a la toma de decisiones.

2.1.1 Modelos relacionales Data Warehouse

Una de las formas utilizadas para la representación de un Data Warehouse, son los modelos relacionales. En ellos existen dos tipos de elementos clave, las tablas de dimensiones y las tablas de hechos. Las tablas de dimensiones representan los factores por los cuales se analiza un determinado objetivo o área del negocio, por ejemplo localización, artículo, etc. Las tablas de hechos están relacionadas con las tablas dimensionales y son los objetos de los análisis, es decir, las que poseen los datos que permiten llevar a cabo los análisis.

A continuación se verán en profundidad dos tipos de modelos relacionales: Modelo estrella y Modelo copo de nieve.

Modelo Estrella

En el modelo estrella [2], como se aprecia en la Fig. 2.1, suele existir una tabla central sobre la cuál se suelen hacer los análisis (tabla de hechos), y una serie de tablas a su alrededor relacionadas con la central (tablas dimensionales o dimensiones).

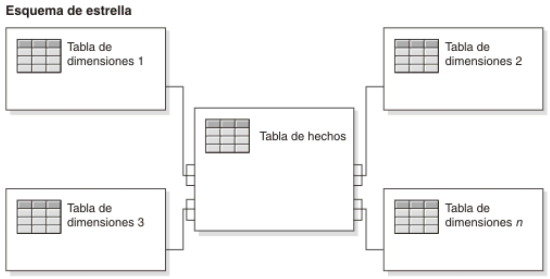

Figura 2.1: Ejemplo modelo estrella (fuente: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-star)

Las tablas de dimensiones siempre tendrán una clave primaria simple y, por otro lado, la tabla de hechos tendrá una clave primaria que está formada por las claves primarias de las tablas de dimensiones.

El nombre del modelo (modelo estrella) viene de la forma en la que la tabla de hechos (de mayor tamaño) es rodeada por las demás tablas (de menor tamaño), dando lugar a una especie de estrella.

Modelo Copo de Nieve

En ocasiones, algunas dimensiones pueden estar compuestas por más de una tabla de datos, por ejemplo, una provincia y sus áreas sanitarias. Dicho de otra forma, lo que se realiza es una normalización de los datos.

Cuando esto sucede, normalmente se trata de un modelo copo de nieve [3], algo más complejo que el modelo estrella, pero es una solución que permite que el espacio de almacenamiento sea más reducido ya que no existe tanta redundancia de datos. La estructura de este modelo se puede apreciar en la Fig. 2.2.

Al existir más tablas de dimensión, a la hora de realizar ciertas consultas, el número de JOINs aumentará y puede empeorar el rendimiento.

2.2 Proceso ETL

Los almacenes de datos obtienen sus datos de diferentes fuentes y éstas pueden ser de distinta naturaleza (como bases de datos, archivos, etc).

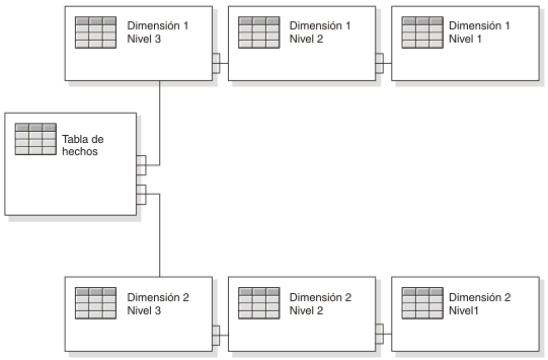

Figura 2.2: Ejemplo modelo copo de nieve (fuente: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-snowflake)

Cuando esto sucede, hay que llevar a cabo un proceso que permita aglutinar en un Data Warehouse los datos necesarios, adecuándolos y realizando las transformaciones necesarias de forma que no existan inconsistencias, datos duplicados, formatos de datos incorrectos, etc.

2.2.1 Pasos ETL

A este proceso se le denomina proceso ETL [4] por sus siglas en inglés *Extract Transform and Load.*

Como bien indica su nombre, se basa en tres fases:

- **Extracción**: Consiste en obtener datos de las diferentes fuentes de datos para poder trabajar con ellos.
- **Transformación**: En esta fase, se estudia el dato que se ha obtenido en el paso anterior, identificando posibles problemas de tipo, nombrado, formato, etc. Es una parte importante ya que se encargará de que los datos de diferentes orígenes sean compatibles entre sí haciendo las modificaciones necesarias para lograrlo.
- **Carga**: Se realiza la carga al Data Warehouse de los datos obtenidos de diferentes orígenes y posteriormente transformados. Gracias a este paso se tendrán los datos disponibles en el Data Warehouse para hacer los estudios que se deseen.

2.3 Análisis de Datos

El análisis de datos [5] consiste en examinar conjuntos de datos con el fin de poder sacar conclusiones, ayudar a tomar mejores decisiones o únicamente obtener información nueva relacionada con los datos analizados.

Normalmente, a la hora de realizar un estudio, se suelen utilizar los datos para buscar evidencias sobre temas relacionados con los mismos, contrastar conocimientos o ampliarlos, descubrir información nueva que se encontraba oculta a simple vista, demostrar sospechas o afirmaciones hechas (o bien refutarlas), etc.

2.3.1 Tipos de Análisis de Datos

Dependiendo del objetivo que se persigue, existen diferentes tipos de análisis, entre los cuales se encuentran los siguientes:

- **Descriptivo**: Tipo de análisis que consiste en sintetizar datos históricos con el fin de obtener información útil. Normalmente son utilizados para resumir datos históricos con el objetivo de encontrar patrones o significados. Sus hallazgos permiten saber si algo ha sucedido o no, o si algo está bien o mal, pero no el por qué. Gracias a este tipo de análisis, se pueden llegar a la formulación de hipótesis.
- **De Diagnóstico**: Este tipo de análisis, al contrario que los descriptivos, no solo buscan saber si algo ha ocurrido, también buscan saber por qué ha ocurrido. Suelen usarse para llegar a conclusiones o resolver cuestiones en las que no basta mirar únicamente los datos ya existentes.
- **Predictivo**: Análisis que tienen como fin descubrir qué es lo que sucederá en el futuro utilizando los hallazgos hechos en los análisis descriptivos y de diagnósticos. Gracias a los datos históricos, es posible predecir eventos que ocurrirán en el futuro.
- **Prescriptivo**: El fin de este tipo de análisis consiste en descubrir qué acciones se deben tomar para que un problema futuro no llegue a ocurrir o aprovechar una tendencia prometedora. Es decir, consiste en intentar averiguar qué acciones llevar a cabo para conseguir los resultados que se deseen.

En este proyecto se hará un análisis descriptivo que permita saber cómo ha afectado el COVID-19 a la población gallega y poder confirmar o desmentir ciertas hipótesis. Sin embargo, en la parte del análisis referente a las restricciones se intentará averiguar si alguna de ellas ha podido propiciar subidas o bajadas de contagios, por lo que también se podría considerar a esta parte un análisis de diagnóstico que busque el por qué sucedieron esos aumentos o descensos de casos positivos.

2.3.2 Ventajas

Los estudios de datos son usados cada vez más debido a que traen consigo una serie de características o ventajas interesantes, algunas de ellas son las siguientes:

- Permite que se tomen decisiones de forma más rápida y documentada ya que están respaldadas por los datos.
- Muestra los datos de una forma más visual, facilitando su comprensión.
- Desde el punto de vista de una empresa, puede proporcionar una ventaja con respecto a las demás (la competencia).

Capítulo 3

Metodologías, Fundamentos tecnológicos y Planificación

3.1 Metodologías

Para poder llevar a cabo el proyecto de forma ordenada y con unas pautas bien marcadas, se hace uso de dos metodologías: Kimball y CRISP-DM. Estas dos metodologías se adaptan a los objetivos que persigue el proyecto. De esta forma, se usan ambas como punto de partida y buena práctica para conseguir un estudio lo más correcto posible.

A continuación se explican detenidamente cada una de ellas y los pasos o fases que las conforman:

3.1.1 Kimball (Diseño de Data Warehouse)

Kimball [6] es una metodología basada en el ciclo de vida dimensional del negocio que se suele emplear para la construcción y explotación de Data Warehouses. El ciclo de vida de un proyecto Data Warehouse [7] se basa en cuatro principios:

Se centra en el negocio, usuario final, cliente...; construir una infraestructura de información adecuada; entregas de forma incremental con avances significativos en cada una; ofrecer una solución completa que tenga todos los elementos necesarios y que aporten valor al usuario final.

Como el diseño, creación y explotación de un Data Warehouse es una tarea difícil de abordar, se utiliza esta metodología para ayudar a simplificar el proceso siguiendo una serie de pasos:

Planificación de proyectos

Se determina cuáles son los propósitos del proyecto Data Warehouse, cuáles son sus objetivos, alcance, riesgos e incluso una aproximación a las posibles necesidades de información que pudiese tener.

Definición de Requerimientos del negocio

Suele ser un proceso en el que se realizan entrevistas para recopilar y ampliar la información previa que se tenga.

La finalidad de este proceso es interpretar de forma correcta los diferentes niveles de requerimientos de los usuarios para poder comprender cuáles son los factores claves del negocio (en este caso es el propio alumno) y poder traducirlo en un modelo apropiado.

Modelo Dimensional

Gracias al paso anterior se obtiene la información necesaria para poder crear un modelo dimensional de alto nivel.

Este proceso suele caracterizarse por ser muy iterativo y se divide en cuatro pasos:

1. Elegir el proceso de negocio: La dirección, tras analizar detenidamente los requerimientos y los temas analíticos del anterior paso, decide qué área modelizar.

2. Establecer el nivel de granularidad: Se especifica el nivel de detalle, el cual se decide en base a los requerimientos de negocio y las posibilidades que hay con los datos actuales.

3. Elegir las dimensiones: Las dimensiones están compuestas por atributos que aportan una perspectiva sobre una medida de las tablas de hechos. Una de las formas de identificar dimensiones es fijarse en sus atributos, estos deben poder ser utilizados como encabezado de informes.

4. Identificar medidas y las tablas de hechos: Se identifican las medidas que surgen de los procesos de negocio. Una medida puede definirse como un campo de una tabla que se quiere analizar usando los criterios de corte (dimensiones).

Diseño Físico

El diseño físico se basa en seleccionar las estructuras que sean necesarias para soportar el diseño lógico de la etapa previa teniendo en cuenta los requerimientos impuestos y los estándares que tendrá el propio Data Warehouse.

Diseño e implementación del subsistema de ETL

El Data Warehouse obtiene sus datos de diferentes fuentes y, para cargar dichos datos en él, primero deben pasar el proceso ETL. El hecho de diseñar este proceso de una forma adecuada permite que se extraigan los datos de los sistemas de origen aplicando una serie de reglas que proporcionen una mayor calidad y consistencia de los datos. Además permite consolidar la información que proviene de distintos sistemas y, por último, cargar la información en el Data Warehouse en un formato adecuado para que pueda ser utilizada por las herramientas de análisis.

Implementación

En esta fase convergen tres aspectos: la tecnología, los datos y las aplicaciones de usuario finales. Además, también existen factores extras (soporte técnico, estrategias de feedback, etc.) que permiten su correcto funcionamiento.

Mantenimiento y Crecimiento de Data Warehouse

Como los usuarios de negocio son el principal motivo por el que existe el propio Data Warehouse, es vital centrarse en ellos para administrar correctamente el entorno del almacén de datos. Además, también es importante gestionar correctamente las operaciones del Data Warehouse, medir y proyectar su éxito y comunicarse constantemente con los usuarios para poder establecer un flujo de retroalimentación. En esto se basa el mantenimiento de un almacén de datos.

Especificación de aplicaciones de BI

Se proporciona una forma más fácil de acceder al Data Warehouse. Esto se hace a través de aplicaciones de inteligencia de negocio o aplicaciones de BI. Éstas son lo que se llama cara visible de la inteligencia de negocio o BI, es decir, los informes o las aplicaciones de análisis que proporcionan información útil a los usuarios.

Las apliacciones de BI comprenden varios tipos de informes y herramientas de análisis que permiten realizar desde informes muy simples, hasta aplicaciones analíticas con algoritmos muy complejos. Kimball permite dividir estas aplicaciones en dos grupos:

Informes estándar: Hacen referencia a informes simples. Proporcionan al usuario final información básica sobre lo que está sucediendo en un área en concreto de la empresa. Por norma general se utilizan diariamente.

Aplicaciones analíticas: Tienen una mayor complejidad que los informes estándar. Pueden incluir algoritmos y modelos de minería de datos, gracias a los cuales se pueden identificar elementos subyacentes a los datos.

Diseño de la Arquitectura Técnica

El área de la arquitectura técnica comprende los procesos y herramientas que son aplicadas a los datos.

En el área técnica existen dos conjuntos, pero sus requerimientos son muy diferentes. Estos dos conjuntos son back room y front room y, debido a sus diferencias, son considerados por separado.

Mientras que el encargado de obtener y preparar datos es el back room, el front room se responsabiliza de entregar los datos a los usuarios.

3.1.2 CRISP-DM (Minería de datos)

CRISP-DM (Cross Industry Standard Process for Data Mining) [8] se ha convertido en una metodología muy popular para aplicar a la minería de datos. Ésta surge de las empresas DaimlerChrysler y SPSS, dos empresas de renombre dentro del mundo de la minería de datos.

Esta metodología establece un proyecto de minería de datos como una secuencia de varias fases:

Comprensión del negocio

Es necesario que los objetivos del proyecto de data mining y los del negocio sean complementarios y persigan el mismo fin, de esta forma los resultados de la minería de datos tendrán un impacto positivo y útil para la organización.

Esta primera fase tiene que englobar los siguientes temas: Establecer los objetivos de negocio, evaluar cuál es la situación actual, fijar los objetivos a nivel minería de datos y obtener un plan de proyecto.

Comprensión de los datos

Este paso se centra no solo en conocer los datos, pero también también su estructura, distribución y calidad.

Así pues, englobará: La ejecución de procesos de captura de datos, poder proporcionar una descripción del conjunto de datos, la realización de tareas de exploración de datos y la gestión de la calidad de los propios datos, identificando problemas y aportando posibles soluciones.

Preparación de datos

Tiene como fin obtener los datos finales sobre los cuales se aplicarán los modelos

En esta fase se persigue: Establecer el universo de datos con los que se va a trabajar, realizar las tareas de limpieza de datos, construir un conjunto de datos para ser usado en modelos de minería de datos y, por último, integrar dichos datos de fuentes heterogéneas si es necesario.

Modelado

Se basa en realizar la construcción de un modelo que sea capaz de alcanzar o satisfacer los objetivos del proyecto en cuestión.

En este caso, se intentan alcanzar los siguientes objetivos: Seleccionar las técnicas de modelado que más se adecuen al conjunto de datos y objetivos, fijar una estrategia de verificación de la calidad del modelo, construir un modelo a partir de la aplicación de las técnicas que se han seleccionado sobre el conjunto de datos y, para finalizar, ajustar el modelo evaluando tanto su fiabilidad como su impacto en los objetivos fijados.

Evaluación del modelo

En esta fase se busca saber cómo de cerca está el modelo creado de cumplir con las expectativas y alcanzar objetivos de negocio marcados anteriormente.

Se debe poder: Evaluar el modelo o modelos generados hasta el momento, revisar todo el proceso de minería de datos hasta el momento, establecer cuales son los siguientes pasos a tomar (mejorar algún paso anterior, seguir con nuevas líneas de investigación, etc.)

Despliegue

Finalmente, una vez obtenidos los resultados, estos se despliegan de forma que se propagan a los usuarios finales. Una vez los usuarios finales ya disponen de los resultados, se procede con el mantenimiento.

Lo que se debe conseguir en esta fase es: Diseñar un plan de despliegue de modelos y conocimientos sobre la organización, realizar el seguimiento y mantenimiento de la parte más operativa del despliegue y revisar el proyecto en su globalidad para poder identificar lecciones aprendidas.

3.2 Fundamento tecnológico

Para la realización de este proyecto, se han utilizado varias herramientas y tecnologías.

3.2.1 PostgreSQL

PostgreSQL [9] es un Sistema de Gestión de Bases de Datos. Éste dispone, en concreto, de una herramienta llamada pgAdmin, la cual permite administrar Bases de Datos de una forma muy fácil e intuitiva gracias a su interfaz gráfica.

Esta herramienta ha permitido crear la base de datos en la que se recogen todas las tablas y, por ende, los datos utilizados para la realización de análisis de los mismos.

3.2.2 Pentaho

Pentaho [10] se trata de una herramienta utilizada para BI. Con ella se pueden llevar a cabo procesos ETL (Extracción, Transformación y Carga). Este proceso permite que datos que pueden venir de diferentes fuentes sean compatibles y adecuados para ser utilizados a la hora de realizar informes que permitan, por ejemplo, asimilar mejor la información, tomar mejores decisiones, etc.

Pentaho es un software que engloba varias herramientas, entre ellas se encuentra Pentaho Data Integration que es la que se ha utilizado en este trabajo para la realización del proceso ETL de cada tabla.

3.2.3 Power BI

Power BI [11] es una herramienta utilizada para la generación de informes.

A pesar de que normalmente se hace mención a Power BI refiriéndose a la aplicación de escritorio (Power BI Desktop), se trata de un conjunto de aplicaciones y servicios que están basados en la nube.

Tiene como finalidad ayudar a recopilar, gestionar y analizar datos que pueden provenir de diversas fuentes. Todo ello puede hacerse de forma sencilla a través de su interfaz intuitiva.

Se encarga de reunir los datos y procesarlos de tal manera que los transforma en información que puede representarse de varias formas (tablas, diferentes tipos de gráficos...) que permitan una mejor percepción de la misma.

El objetivo es que, mediante esta forma de representar la información, se facilite a los usuarios la posibilidad de realizar informes y cuadros de mando útiles que ayuden a tomar decisiones más acertadas y estudiadas.

3.3 Planificación

A la hora de llevar a cabo un trabajo de estas magnitudes, lo que se debe hacer es realizar una planificación inicial para abordar de la mejor manera posible el proyecto.

De esta forma, se procede con la creación de un diagrama de Gantt, que permitirá reflejar esta planificación de una forma más visual haciendo que sea sencilla de comprender y seguir. En él se recogerán las diferentes etapas o tareas que se deberán llevar a cabo desde el inicio hasta el fin del proyecto, indicando una supuesta fecha de inicio y fin para cada una de ellas.

El número de días asignados a cada tarea se basa la dificultad de la propia tarea y el tiempo que se pueda dedicar a ella los días que tiene asignados.

Como resultado de la planificación, se obtiene el diagrama de Gantt, que se puede ver en la Fig. 3.1.

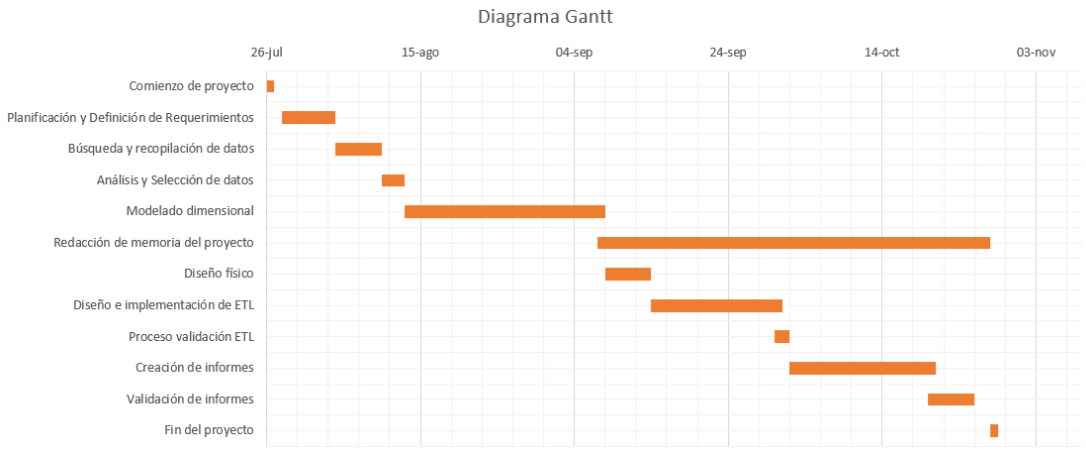

Figura 3.1: Diagrama que representa la evolución que se desea que tenga el proyecto a lo largo del tiempo.

A pesar de que la planificación se intenta hacer de la manera más exacta posible, es normal que haya diferencias entre lo planeado y lo real. Estas desviaciones de la planificación pueden suceder por problemas nuevos que aparecen y dificultan una tarea, o bien el caso contrario, en el que una tarea se finaliza antes de lo esperado. Como resultado de estas desviaciones, al finalizar el proyecto se elabora un nuevo diagrama de Gantt con la planificación que finalmente se ha llevado a cabo, la cual no debería ser muy diferente a la inicial.

En este nuevo diagrama, que se puede ver en la Fig. 3.2, se precian pequeñas diferencias de tiempos en algunas tareas (con respecto al anterior diagrama). Las más notorias son durante la tarea de búsqueda y recopilación de datos y las del diseño, creación y validación del proceso ETL. Esto es debido a que a la hora de recoger los datos de las restricciones, éstos venían a modo de redacción en formato PDF. En cuanto al diseño y creación de ETL hubo ciertos problemas con la configuración del programa y hubo que repetir algún proceso que inicialmente estaba tratando de forma errónea el dato y hasta la fase de validación no se percibió. La validación de los procesos ETL es el caso contrario, este ha llevado menos tiempo del esperado ya que la propia validación se hacía más rápido de lo que se suponía.

Además la creación de informes fue una tarea más sencilla y rápida de lo esperado debido al conocimiento más profundo de los datos y a tener muy claro lo que se quería mostrar en dichos informes.

Por último, la elaboración de la memoria parecía que llevaría menos tiempo pero finalmente, a la hora de pulir pequeños errores y cambiar algunas expresiones, se alargó el tiempo de la tarea.

También es interesante destacar que la planificación inicial estaba pensada una jornada diaria de una hora en las fases tempranas y dos horas en las restantes debido a agentes externos como la jornada laboral del alumno. A pesar de haberlo planeado así, el resultado fue muy

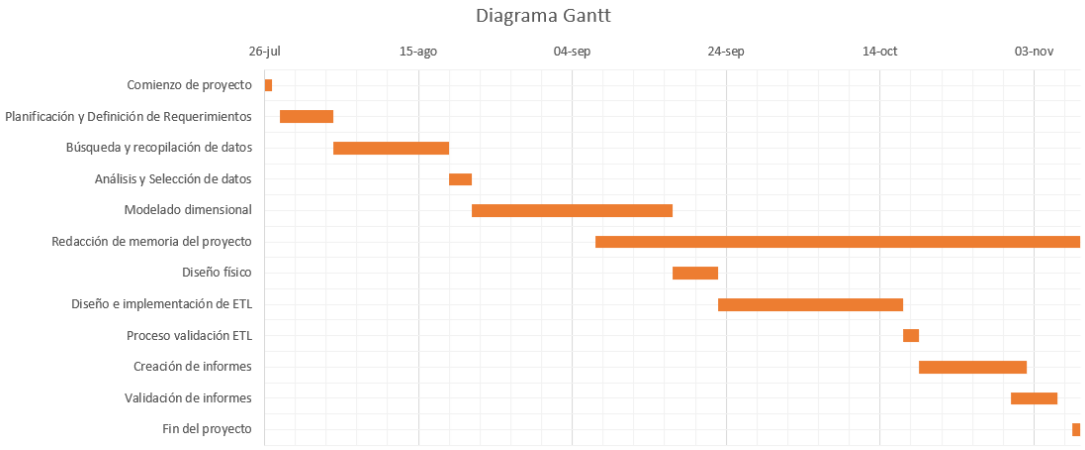

Figura 3.2: Diagrama que representa la evolución real del proyecto a lo largo del tiempo.

distinto ya que al principio sí que se respetó esa hora diaria pero a medida que iban avanzando los días, las dos horas diarias aumentaban considerablemente (sobre todo las últimas semanas). Por eso, mientras que en un principio se había pensado que el proyecto durase sobre 120 horas, se ha acabado convirtiendo en un proyecto de, más o menos, 230 horas. Si se pusiese un coste de 25 €/hora (descontando el tiempo de la realización de la memoria) se trata de un proyecto que ronda un coste de 4.700€.

Capítulo 4

Análisis y Diseño del Data Warehouse

En este capítulo se abordarán, detalladamente, los primeros pasos que conducen al análisis de datos sobre el COVID-19 en Galicia y las restricciones impuestas en sus provincias. Estos pasos comprenden desde el estudio de los datos que se desean tener en el Data Warehouse y su búsqueda, hasta el diseño del mismo.

4.1 Análisis del Data Warehouse

El primer paso a la hora de realizar un análisis de datos es determinar cuáles son los requisitos y los objetivos que se desean alcanzar. Una vez hecho esto, se identifican los datos necesarios para poder llevar a cabo el análisis y alcanzar el objetivo del estudio.

Hoy en día la abundancia de datos es un aspecto tanto positivo como negativo. El poder obtener muchos y variados datos es algo bueno ya que permite obtener prácticamente cualquier información necesaria para el estudio, pero también puede ser malo. El exceso de información puede llegar a confundir y seguramente se tenga información redundante, innecesaria, etc.

Así pues, es muy importante decidir qué datos se necesitan y cuáles son interesantes para cumplir el objetivo del análisis.

Para recolectar los datos usados para este proyecto, se formula la siguiente pregunta:

¿Qué es lo que se quiere demostrar con este estudio?

O, dicho de otra forma: ¿Cuál es el fin que se persigue?

La respuesta a ambas preguntas converge en poder conocer cuáles fueron las consecuencias reales del COVID-19 en las provincias y/o áreas sanitarias gallegas (número de contagios, muertes, hospitalizaciones, personas más contagiadas, etc.). Además de ello, también se busca conocer la repercusión de ciertas restricciones en las diferentes áreas sanitarias de Galicia para poder determinar cuáles tuvieron más o menos impacto, si esta fue positiva o negativa

para la población, etc.

De esta forma, se comenzó con la búsqueda de datos sobre la población gallega, priorizando, entre otros, los siguientes campos:

- Sexo de los individuos.
- Grupo edad de los individuos.
- Número de muertes diarias.
- Número de contagios o casos confirmados de COVID-19.
- Número de pacientes hospitalizados.
- Las fechas en las que las restricciones estuvieron activas en cada zona.
- Número de pruebas realizadas por zona.

Todos ellos se priorizan ya que son los necesarios para poder alcanzar el objetivo del estudio. Gracias a estos datos se podría determinar si Galicia tuvo muchos o pocos casos positivos, la letalidad del virus (la cantidad de individuos que mueren tras contraer la enfermedad), saber los ingresos en hospitales producidos por el COVID-19, etc.

Además, si se tuviesen estos datos agrupados por edades y sexo, el estudio sería mucho más rico ya que se podría determinar, por ejemplo, qué edades son las más afectadas o si el sexo del individuo es relevante para saber si tiene más o menos posibilidades de fallecer.

Si también se poseen datos de las restricciones activas por fechas, combinando estos con, por ejemplo, el número de contagios, se haría posible el poder realizar investigaciones para determinar qué restricción causó una caída más drástica de contagios, cuál supuso una subida, si se sigue cierto patrón a la hora de implantar una restricción, etc.

Es importante recalcar que todos los datos estén a nivel de provincia o área sanitaria gallega para poder realizar el estudio.

4.2 Búsqueda de datos

Al saber exactamente qué datos se necesitan, se puede enfocar la búsqueda de éstos y hacerla de una forma más eficiente.

Una de las páginas utilizadas para la recolecta de datos es la propia web del SERGAS [12] (en gallego, Servizo Galego de Saúde) que se encarga de la asistencia sanitaria pública de Galicia.

En ella se pueden encontrar multitud de datos referentes al COVID-19. En concreto existe un apartado en el que se muestran (a nivel de área sanitaria) datos generales en el día actual

sobre la enfermedad, lo cual los hacen idóneos para la investigación. El problema en este caso es que únicamente muestra datos del día anterior, por lo que no dispone de datos de los días previos. De esta forma se procede con la búsqueda de alguna página o repositorio en el que se hubiesen ido recolectando datos desde fechas anteriores, no solo la del día previo.

La búsqueda fue fructífera ya que existe un repositorio en el que se recogen esos datos (los que facilita la página web del SERGAS) desde octubre del año 2020 [13].

De esta forma, se puede hacer un análisis a nivel área sanitaria que permite conocer el número de fallecidos, contagios, pruebas PCR hechas, etc.

A pesar de haber encontrado datos muy útiles, se echan en falta otros campos como los grupos de edades y sexo, los cuales serían de gran interés a la hora de realizar un análisis del impacto del COVID-19 en cada grupo de individuos.

Como el análisis se centra también en las provincias de Galicia (no solo áreas sanitarias), se comienza a buscar información a un nivel mayor, es decir, en lugar de centrar la búsqueda en páginas gallegas, ésta amplía su rango y se incluyen páginas que lleven recuentos a nivel provincias (de España). Lo que se consigue de esta forma es que la búsqueda devuelva más resultados y, por consecuencia, haya más páginas con datos de la pandemia. Es posible que en estas nuevas páginas los datos estén agrupados por edades y sexo. Además, se podría hacer algún pequeño análisis a nivel nacional a mayores de las provincias de Galicia.

Finalmente, con esta búsqueda se obtienen datos a nivel de provincia que aportan datos de edades y sexo, es decir, justo lo que se perseguía. Esta información fue sacada de la página [14], otro repositorio en el que se recoge información sobre la pandemia, pero en este caso a nivel provincia.

El último paso es encontrar información referente a las restricciones, para ello se hace uso de la página oficial de la Xunta de Galicia, en la que se pueden encontrar numerosos ficheros PDF que redactan algunas de las restricciones que se implantan en cada zona de la comunidad autónoma.

Este punto es el más engorroso a la hora de recolectar datos ya que no se trata de obtener directamente un archivo CSV, la información se recoge de varios PDFs y, tras analizarlos, se rellena manualmente un archivo CSV que recoge la información en el formato correcto para poder llevar a cabo el análisis.

De esta forma se crea un nuevo archivo CSV con datos sobre las restricciones en Galicia que, a priori, no se encuentra en internet, al menos en las páginas oficiales de sanidad.

Con él se podrán realizar estudios sobre las restricciones como, por ejemplo, saber si el hecho de quitar o imponer una restricción implica un aumento o descenso de contagios o muertes.

Sobre esta parte de restricciones, es importante dejar claro que, por el alcance que tiene este proyecto, no se vio viable el hecho de recoger el dato a nivel de ayuntamientos de Galicia.

Esto significaría obtener, por cada uno de los 313 municipios de Galicia, más de 4.000 registros, algo inviable para el alcance de este estudio.

Por esta razón, lo que se hizo fue escoger un grupo de municipios de las áreas sanitarias teniendo en cuenta cuáles eran los más representativos, los más grandes, en los que había más contagios, etc. De esta forma, observando qué restricciones estuvieron activas en ese grupo de municipios a lo largo del tiempo, se determinan las restricciones activas en cada área sanitaria durante la pandemia.

En el caso de querer realizar el estudio a un nivel superior de área sanitaria, por ejemplo provincia, el proceso sería el equivalente al anterior. Así pues, para determinar si una provincia tiene una restricción activa en una fecha determinada, se procede del mismo modo, pero esta vez utilizando las áreas sanitarias más representativas de cada provincia (más contagios, habitantes, etc.) en lugar de los ayuntamientos (o municipios de cada área sanitaria).

Por todo esto, se deja claro que la parte del estudio basada en las restricciones se basa en una estimación de las restricciones de cada provincia y/o área sanitaria gallega. Es decir, se estima que toda la provincia (o área sanitaria) comparte las mismas restricciones (en las mismas fechas) que sus áreas sanitarias (o municipios) más importantes y/o representativas (ya sea por significado, población, contagios, etc.).

Por último, ya que el estudio tratará sobre las provincias y áreas sanitarias, se recoge información sobre ellas, buscando campos que puedan ser de utilidad para realizar agrupaciones a la hora de generar los informes, como población, densidad…

4.3 Diseño Data Warehouse

4.3.1 Modelo Conceptual

Tras identificar todos los datos que se podrían utilizar en el Data Warehouse, se elabora un modelo conceptual Fig. 4.1 del propio almacén de datos. En él se realiza una descripción a alto nivel del contenido que tendrá el Data Warehouse independientemente del Sistema de Gestión de Bases de Datos (SGDB) que se vaya a utilizar. No solo se indican los datos que se almacenarán, también se muestra la forma en la que estos se relacionan entre sí a un alto nivel. Gracias a este modelo conceptual se pueden distinguir también cuáles serán los hechos y dimensiones.

De esta forma, como se puede apreciar en la Fig. 4.1 , nos encontramos con tres hechos y cinco dimensiones, cada uno de ellos formados por diferentes atributos o campos. A continuación se hablará de ellas y de sus elementos de una forma detallada.

Dimensión de área sanitaria

Posee la información referente a cada área sanitaria de Galicia. Recoge campos que permitan realizar agrupaciones a la hora de realizar los informes. En ella se encontrarán los siguientes campos:

- **area_sanitaria**: Hace referencia al nombre por el que se conoce al área sanitaria y permite identificarla unívocamente.
- **provincia**: Indica la provincia a la que pertenece el área sanitaria.
- **población**: Representa el número de habitantes (atributo cambiante).
- **zona**: Permite conocer la parte de Galicia a la que pertenece el área sanitaria (Norte, Sur, Este, Oeste y sus combinaciones).
- **es_unica_en_provincia**: Usada para saber si hay más áreas sanitarias en la provincia a la que pertenece.
- **nivel_poblacion**: Indica a qué grupo pertenece el área sanitaria desde el punto de vista de la población. Puede tomar los siguientes valores: bajo (<150.000), medio (>=150.000 y <250.000), alto (>=250.000).

La población de un área sanitaria puede variar a lo largo del tiempo, por lo que se trata de una dimensión cambiante. Al tratarse de una dimensión cambiante se necesita modelarla de tal forma que se puedan gestionar los cambio (se hará en la sección 4.3.2).

Dimensión de provincia

Siguiendo la misma dinámica que la anterior dimensión, ésta poseerá los datos de las diferentes provincias de España. Los campos que conforman la dimensión son los siguientes:

- **provincia**: Campo que hace referencia al nombre de la provincia y permite identificarla unívocamente.
- **provincia_iso**: Es un código que identifica cada provincia de España.
- **población**: Número de habitantes en la provincia.
- **nivel_población**: Indica si la provincia tiene un nivel de población bajo (<350.000), medio (>=350.000 y <800.000) o alto (>=800.000 habitantes).
- **%poblacion_españa**: Indica el porcentaje de población que representa dicha población en todo el país.

- **nivel_%poblacion_españa**: Dependiendo del porcentaje de población de cada provincia, indica su nivel: bajo (<0.75), medio (>=0.75 y <1.70) o alto (>=1.70).
- **densidad(Hab/Km2)**: Muestra la densidad de cada provincia, resultante de la división del número de habitantes entre los kilómetros cuadrados de la provincia.
- **nivel_densidad**: Dependiendo del valor de densidad de la provincia, ésta pertenecerá a un nivel de densidad bajo (<30), medio (>=30 y <130) o alto (>=130).

Al igual que el caso anterior, esta dimensión también es cambiante ya que el campo población (y los que dependen de él) puede variar a lo largo del tiempo y por ello se debe modelar la dimensión para que pueda gestionar estos cambios (se hará en la sección 4.3.2).

Dimensión de restricción

Dimensión que contiene información sobre las restricciones que se han utilizado en Galicia durante la pandemia.

- **id_restriccion**: Campo que permite identificar unívocamente una restricción.
- **descripción**: Breve descripción de la restricción.
- **nivel**: Indica el nivel a partir del cuál se empezará a usar dicha restricción. Por ejemplo, el uso de mascarilla, tendrá un nivel bajo debido a que a pesar de que los contagios bajen estrepitosamente, es necesario seguir usándola. Por el contrario, un confinamiento absoluto, sería de nivel alto ya que es una restricción usada en situaciones con muchos contagios.
- **tipo_restriccion**: Permite saber a qué tipo pertenece la restricción, un tipo genérico, de ocio, de movilidad o de protección. Para que se entiendan mejor los tipos de restricciones, se facilita a continuación una breve descripción de cada uno:
  - **Tipo Genérico**: Aquellas que afectan a más de un tema, por ejemplo el nivel de restricción alto o el confinamiento (afecta a la movilidad, a la hostelería, socialización...)
  - **Tipo Ocio**: Hace referencia a cierres de lugares de, como su nombre indica, ocio (comerciales, hostelería, lugares de fiestas...)
  - **Tipo Movilidad**: Son aquellas que restringen de alguna forma la posibilidad de desplazamiento libre de los individuos, por ejemplo los cierres perimetrales.
  - **Tipo Protección**: Se refiere a las restricciones que deben realizar los ciudadanos y que no pertenecen a ningún otro tipo.

En este caso, la dimensión no es cambiante, por lo que no será necesario modelarla de forma que se tengan en cuenta cambios en sus registros.

Dimensión de tiempo o fecha

En este caso, se trata de una dimensión que engloba la información referente al tiempo, tanto las fechas como sus características (día de la semana, mes, etc.)

- **fecha**: Representa la fecha completa en formato yyyy-MM-dd. Es usada como clave primaria de la tabla.
- **dia**: Hace referencia al día de la propia fecha. Por ejemplo, fecha->2020-10-20, día->20
- **Mes**: Sigue el mismo estilo que el campo anterior, hace referencia al mese de la fecha. Por ejemplo, fecha->2020-10-20, mes->10
- **ano**: Indica el año de la fecha
- **dia_semana**: Comenzando en domingo y terminando en sábado, representa (con números del uno al siete) los días de la semana.
- **dia_semana_nombre**: Día de la semana en formato texto (Lunes, Martes, ...)
- **mes_nombre**: Mes en formato texto (enero, febrero, ...)

En este caso, como tampoco se trata de una dimensión cambiante, se modela sin tener en cuenta cambios en el tiempo.

Dimensión de tipo de paciente

En ella se recogerá únicamente el sexo y el grupo de edad al que pueden pertenecer los individuos.

- **tipo_paciente**: Campo que permite identificar unívocamente un tipo de paciente.
- **sexo**: Puede tener tres valores, NC, M y H. Éstos hacen referencia a: personas que no se conoce su sexo (bien porque no lo han especificado o porque no se consideran ni mujeres ni hombres), Mujeres y Hombres respectivamente.
- **grupo_edad**: Rango de edad al que pertenece un paciente: NC (no se sabe su edad), 0-9,10-19,...,70-79,+80 (80 años o más)

Esta dimensión, al igual que las dos anteriores, no es cambiante.

Hecho Casos_Generales

En él se encuentran las métricas que permitirán realizar estudios sobre el COVID-19 en las diferentes áreas sanitarias de Galicia. Se relaciona con las dimensiones de tiempo y área sanitaria. Los atributos que lo forman son los siguientes:

- **casos_confirmados_pcr_ultimas_24h**: Número de casos confirmados por pruebas PCR en las últimas 24 horas (métrica aditiva).
- **camas_ocupadas_uci**: Camas ocupadas del área UCI en una fecha (métrica aditiva).
- **camas_ocupadas_hos**: Camas ocupadas sin ser del área UCI en una fecha (métrica aditiva).
- **casos_totales**: Número de casos positivos en COVID-19 en una fecha (métrica aditiva).
- **exitus**: Número de muertes en una fecha (métrica aditiva).
- **pacientes_con_alta**: Número de pacientes a los que se les ha dado el alta en una fecha (métrica aditiva).
- **pacientes_sin_alta**: Número de pacientes a los que no se les ha dado el alta en una fecha (métrica aditiva).
- **pruebas_realizadas_no_pcr**: Número de pruebas diferentes a PCR realizadas en una fecha (métrica aditiva).
- **pruebas_realizadas_pcr**: Número de pruebas PCR realizadas en una fecha (métrica aditiva).

Hecho Prov_Casos

Hecho que recoge las métricas que permitirán hacer estudios sobre el COVID-19 en las diferentes provincias de España. En él se encuentran datos a nivel de día, provincia y tipo de paciente (que indica el sexo y la edad de los individuo).

- **num_casos**: Número de casos positivos en COVID-19 en una fecha (métrica aditiva).
- **num_hos**: Número de personas hospitalizadas en una fecha (métrica aditiva).
- **num_uci**: Número de personas ingresadas en UCI en una fecha (métrica aditiva).
- **num_def**: Número de personas fallecidas en una fecha (métrica aditiva).

Hecho Area_Fecha_Restriccion

Este es un hecho un poco diferente a los anteriores ya que se trata de una factless (no tiene hechos), únicamente permite saber si una restricción estuvo activa a lo largo de la pandemia. Este hechos estaría conformado por:

- **activo**: Indica si una restricción estuvo activa en un área sanitaria y fecha determinadas.

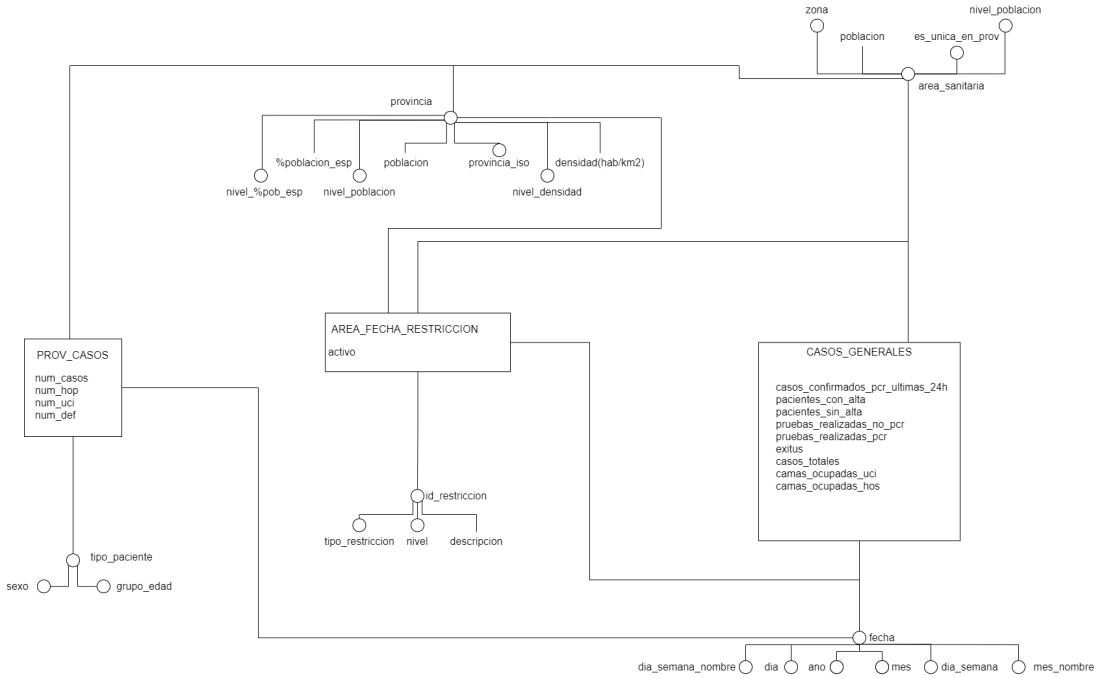

Figura 4.1: DFM Data Warehouse

4.3.2 Modelo Lógico

Tomando como punto de partida el modelo conceptual del apartado anterior, se procede con el modelo lógico Fig. 4.2.

Con él se busca moldear la estructura que tendrá el Data Warehouse, pero esta vez teniendo en cuenta que utilizará un sistema gestor de bases de datos relacional. La finalidad del modelo lógico es obtener una representación que use los recursos disponibles de la manera más eficiente posible para estructurar y modelar las condiciones.

En el modelo lógico de este proyecto se diseñan lo que serán las tablas de hechos y dimensiones, indicando los campos que tendrá cada tabla, su tipo y cómo se gestionan las relaciones entre dichas tablas. Además, se tratan otros temas como puede ser la gestión de cambios planteada en el modelo conceptual.

Este es un punto importante ya que en él se decide el esquema o modelo del Data Warehouse, tratándose en este caso de un modelo estrella (descrito en la sección 2.1.1).

Gracias a las figuras Fig. 4.1 y Fig. 4.2, se puede apreciar que cada hecho y dimensión del modelo conceptual, se transforma en tablas de hechos y de dimensiones respectivamente. Se muestran tanto los campos de las tablas (también indicados en el modelo conceptual) como sus tipos y la forma en la que se gestionan las relaciones.

Gestión de cambios

En esta fase también se decide cómo gestionar los cambios de las dimensiones 'dim_provincia' y 'dim_area_sanitaria'. La solución por la que se ha optado es SCD tipo 2. De esta forma se añade una clave subrogada en cada una una de las dimensiones ('id_provincia' e 'id_area_sanitaria') que identifiquen unívocamente cada registro. De esta forma gracias a la clave subrogada se podrá identificar cada registro y con la clave natural se podrán hacer consultas propias de la entidad, por ejemplo, teniendo en cuenta todos los registros de una sola provincia (si esta tiene varios registros, o lo que es lo mismo. varias versiones) Esta clave subrogada sustituirá a la clave natural de estas dimensiones que se habían comentado en la sección 4.3.1.

Además, en cada una de estas dimensiones se añadirán dos campos a mayores ("fecha_inicio" y "fecha_fin") que indican la fecha en la que comienza a estar activo ese registro y la fecha en la que deja de ser válido.

Estos campos, permiten que se pueda llevar un seguimiento histórico de la situación de las áreas sanitarias y provincias si se precisase en un futuro.

4.4 Creación del Data Warehouse

El siguiente paso ya sería elaborar el propio Data Warehouse (siguiendo las indicaciones del modelo conceptual y lógico). El resultado final es un almacén de datos formado por siete tablas, tres de ellas de hechos y las cinco restantes de dimensión.

En las tablas de hechos, como ya se ha comentado, se recogen los datos a analizar sobre las provincias (prov_casos) y areas sanitarias (casos_generales). La otra tabla de hechos permitirá saber si una restricción estuvo activa en una fecha y área sanitaria concreta (area_fecha_restriccion). Ésta última es de gran ayuda ya que permite que se estudie si el aumento de contagios pudo ser propiciado por dejar de estar activa una restricción o cuáles fueron las restricciones más y menos utilizadas, por ejemplo.

Por otro lado, las tablas dimensionales se utilizarán, entre otras cosas, para enriquecer los informes y poder realizar estudios desde diferentes puntos de vista. Éstas serán provincia, área sanitaria, tiempo, restricción y tipo de paciente.

Al finalizar esta fase, se tendrá disponible el Data Warehouse en el que se comenzarán a

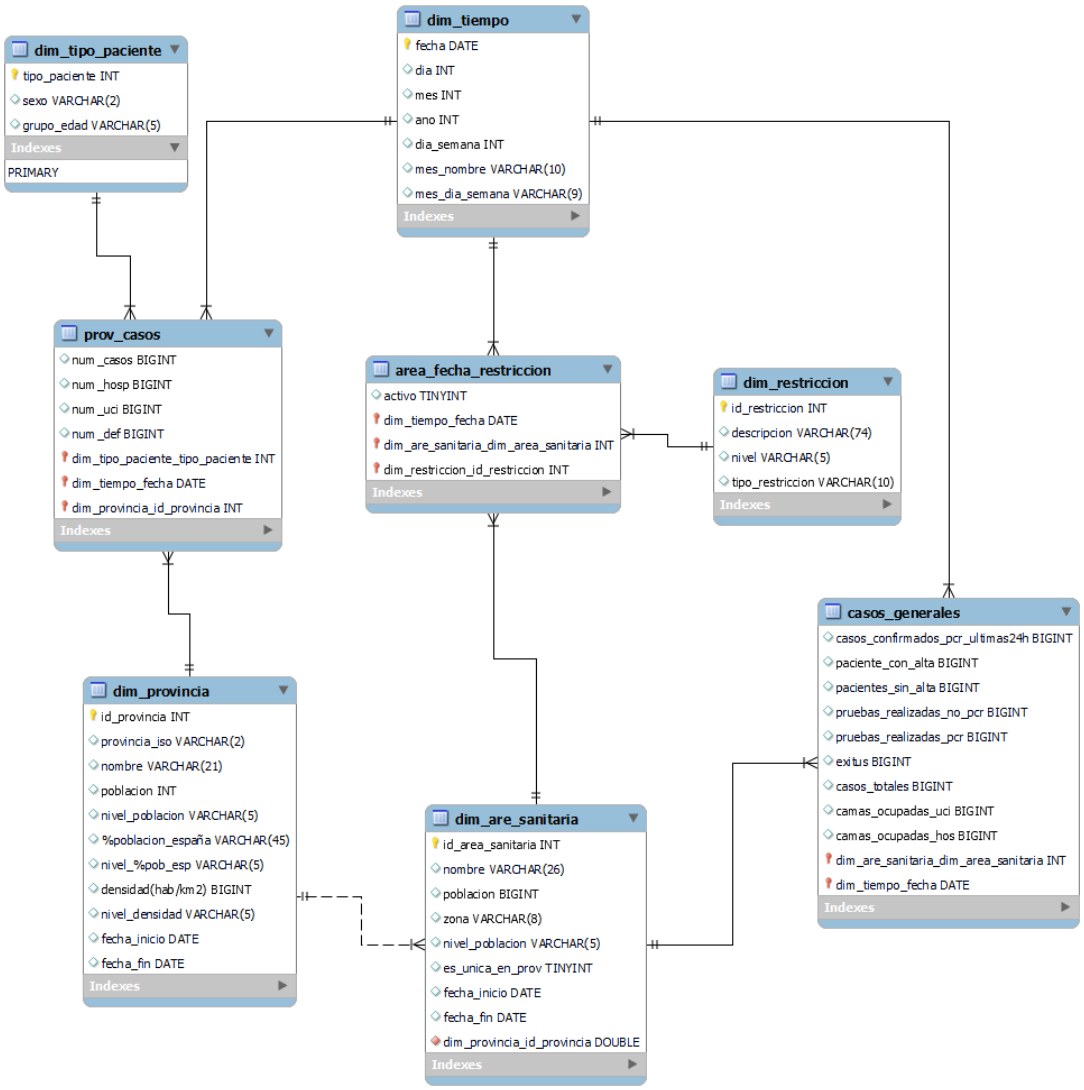

Figura 4.2: Modelo Relacional Data Warehouse

cargar los datos para poder realizar informes con ellos. Este es, por así decirlo, el resultado de la implementación del modelo conceptual y lógico.

Capítulo 5

Proceso ETL

A la hora de completar con datos el Data Warehouse, es muy importante adecuar dichos datos para que no haya ningún error o problema por la forma en la que se almacenan.

Así pues, estos tienen que pasar por un proceso ETL que realiza los cambios necesarios sobre los datos de las diferentes fuentes de origen. Como esto puede ser muy engorroso si se realiza cada vez que se hace una carga, se utiliza una herramienta llamada Pentaho. Pentaho permite leer los datos de las fuentes de origen (en este caso archivos .CSV), realizar las transformaciones de datos que se deseen y cargarlos, en este caso, en tablas del Data Warehouse.

Gracias a esta herramienta, se pueden crear archivos que, al ejecutarlos, realicen el proceso entero, desde la lectura de información de las fuentes de origen hasta la carga de los datos modificados en el Data Warehouse. De esta forma, si se quieren cargar nuevos datos, simplemente se añaden al archivo .CSV que es leído por el archivo de Pentaho y, al ejecutarse este último, se realiza el proceso ETL entero, finalizando con la carga de los datos en el Data Warehouse.

Esta automatización es muy interesante ya que permite sustituir todos los pasos del proceso por la ejecución de un archivo. Desde el punto de vista del administrador del Data Warehouse es una gran ventaja ya que se ahorra muchos pasos en futuras cargas pero también se ha automatizado este proceso pensando en otros posibles usuarios con conocimientos informáticos menos avanzados.

En el caso de que en un futuro se diese la posibilidad de que persona externas al proyecto tuviesen en su equipo el Data Warehouse de este, si se le facilitan los archivos que incluyen el proceso ETL, se le da la posibilidad al usuario de que puedan realizar cargas sin saber cómo funciona por debajo.

En el Data Warehouse de este proyecto existen, en total, cinco tablas de dimensiones y tres tablas de hechos y, para cada una de ellas, se crea un proceso ETL utilizando la herramienta Pentaho. A continuación se explican detalladamente cada uno de ellos, haciendo uso de imágenes para una mejor comprensión.

5.1 DIM_TIEMPO

En primer lugar, la dimensión de tiempo, como se ha comentado es una dimensión particular con respecto a las demás ya que los datos no provienen de fuentes externas, estos se generan en el propio proceso ETL.

Se comienza generando un número de registros a fecha 2020-01-01 que es la fecha en la que se comienza a tener dato para analizar.

A continuación, para poder generar los demás días se crea un contador que va aumentando de uno en uno (hasta el momento solo se tienen varios registros con la misma fecha).

La calculadora nos ofrece una opción que permite sumar un número de días a una fecha y, con ella, se pueden ir sumando los valores del contador a las fechas, de forma que se obtienen todos los días desde el 2020-01-01 en adelante.

Cabe recalcar que a la hora de generar fechas, se le ha indicado a este paso que genere 571 filas concretamente, que cubrirían todos los días desde el 2020-01-01 hasta el 2021-07-24 (fecha donde se deja de tener dato para analizar). En el caso de que, en un futuro se quisiese ampliar este rango de fechas, para no tener que calcular el número de días exacto, se incluye un paso que permite indicar un día límite para que las fechas posteriores a este se descarten y no se carguen en la dimensión Fig. 5.1.

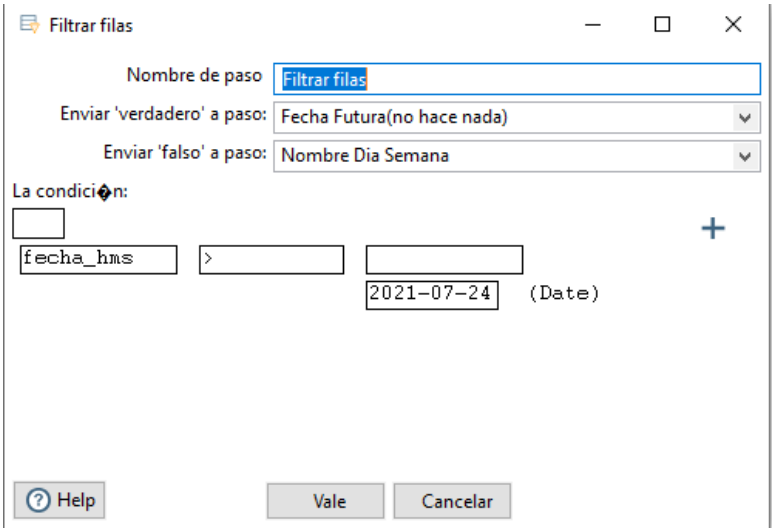

Figura 5.1: Filtro fechas

Una vez obtenidas estas fechas se obtienen, a partir de ellas, otros valores como el día, mes, año y día de la semana. Además, se añaden otros dos campos, que son el nombre del mes y del día de la semana en formato texto, cuyos valores se generan leyendo tanto el mes como el día de la semana en número y se transforman al texto que le corresponde.

Posteriormente, se comparan los datos de la tabla dim_tiempo con los datos que genera

este proceso. Prevalecerán los valores generados en el ETL a los que se encuentran en la propia dimensión. Es decir, en el caso de que un registro esté en la dim_tiempo y no en los generados en el proceso ETL, se marcará como deleted y se eliminará de la dimensión, si el registro simplemente es el mismo pero modificado, se marcará como changed y se modificará (para errores en inserciones), y, si un registro es nuevo (no se encuentra en la dimensión), se marcará como new y se añadirá a la tabla.

El proceso ETL de la dim_tiempo se puede revisar en la Fig. 5.2.

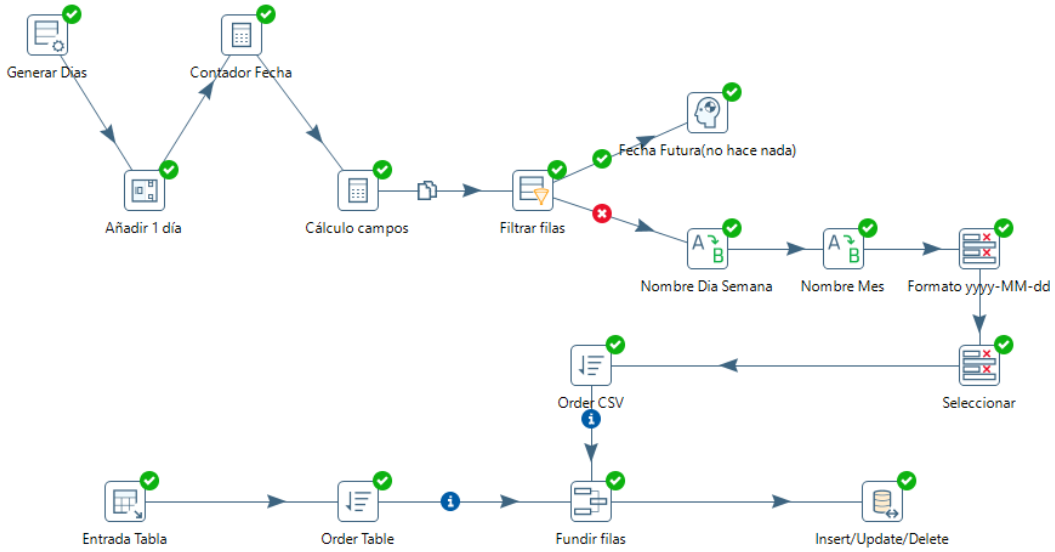

Figura 5.2: Proceso ETL Dimensión Tiempo

5.2 DIM_TIPO_PACIENTE

En la Fig. 5.3 se pueden apreciar los diferentes pasos que se llevan a cabo a lo largo del proceso.

Se comienza leyendo los datos de un archivo CSV el cual debe tener un formato concreto legible para Pentaho (la primera línea es la cabecera, separar los campos con ';' o ',', etc.).

A continuación se generan los identificadores de los registros del archivo CSV y se guardan en un campo nuevo. Este nuevo campo se llama tipo_paciente y no está en el archivo CSV ya que lo genera el propio proceso ETL asegurándose de que no haya duplicados y que cada registro tiene asignado un identificador.

Una vez creado este campo se seleccionan todos y se hacen los cambios o transformaciones necesarias en el tipo de dato de cada campo para que los datos introducidos en el Data Warehouse estén en el formato adecuado.

Por otro lado, se lee también del Data Warehouse la dim_tipo_paciente, con el fin de poder comparar los datos que ya existen en la dimensión con los del archivo CSV.

Se tienen que ordernar los datos de la dimensión y los datos del archivo CSV antes de ser comparados. Tras esto, se utiliza un paso en el que se selecciona un campo que se usará para identificar los registros de cada parte de la comparación y se realizará la comparación sobre los restantes. Este paso devolverá los registros indicando a mayores, en un campo nuevo, si ese registro no ha cambiado (valor del campo 'identical'), si sí lo ha hecho (valor del campo 'changed', pensado para errores de inserción), si hay registros en la dimensión que no están en el CSV (valor del campo 'deleted') o si, por el contrario, hay campos en el CSV que no existen en la dimensión (valor del campo 'new').

Dependiendo del valor del campo en cuestión, se procederá de una u otra forma. En el caso de que se intenten modificar o eliminar registros, no habrá problema (a no ser que las dependencias con otras tablas lo prohíban) pero, en el caso de las inserciones, se debe revisar primero si el registro que se intenta añadir a la dimensión hace referencia a un tipo de paciente ya existente (si coincide su clave candidata 'sexo'-'grupo_edad'). Si este fuese el caso, la inserción no se lleva a cabo y se muestra un mensaje de error indicando el problema, en caso contrario, la inserción se lleva a cabo.

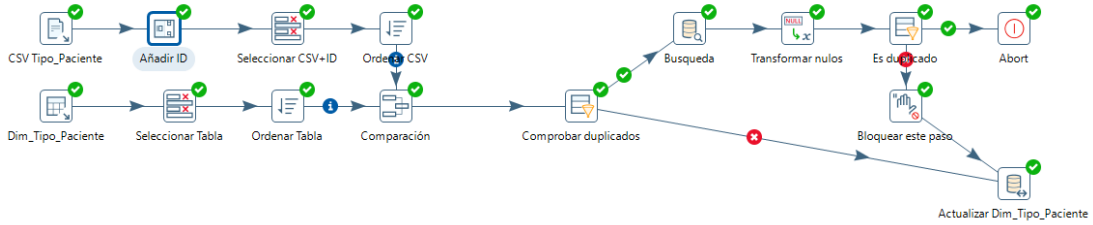

Figura 5.3: Proceso ETL Dimensión Tipo Paciente

5.3 DIM_RESTRICCION

En dim_restriccion se cargan los datos de las restricciones que se analizan en este proyecto.

En este caso, el proceso ETL tiene exactamente los mismos elementos que, la dim_tipo_paciente, situados de la misma forma y en el mismo orden, es decir, son el mismo proceso pero accediendo a un archivo CSV y una tabla del Data Warehouse diferentes y cargando el dato en la dimensión que le corresponde.

La forma de añadir, modificar y eliminar registros de la dimensión también es la misma, se hará todo modificando el CSV y ejcutando el archivo de Pentaho.

En la Fig. 5.4 se puede apreciar el proceso ETL de la dim_restricción y que éste es igual que el del caso anterior, la dim_tipo_paciente.

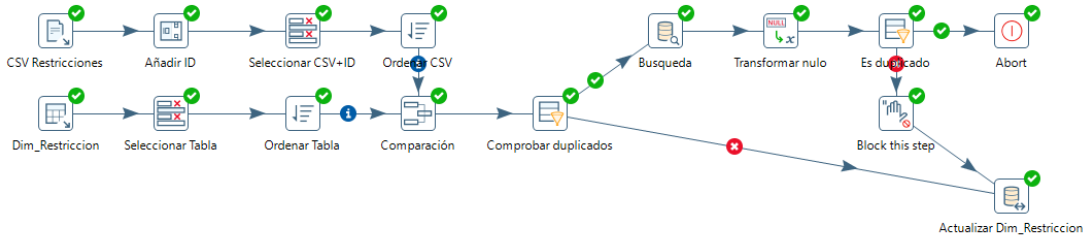

Figura 5.4: Proceso ETL Dimensión Restricción

5.4 DIM_PROVINCIA Y DIM_AREA_SANITARIA

En el caso de las dimensiones de provincia y área sanitaria, se quiere guardar información histórica y, para ello, se usan dos campos que permiten saber la fecha en la que un registro comenzó su período de validez y la fecha que indica el fin de dicho período. Se llevará constancia de esto de forma manual, pero si el proyecto continuase en un futuro, se podría estudiar la posibilidad de automatizarlo. En este caso, al no estar automatizado, cuando se quiere añadir a la dimensión un registro nuevo que hace referencia a una provincia ya existente en la tabla, se debe completar la información del archivo CSV siguiendo las siguientes indicaciones:

- Cambiar el valor del campo fecha_fin del registro previo correspondiente (con la provincia que se desea). El valor debe ser un día antes de la fecha_inicio del nuevo registro. Por ejemplo, si se quiere añadir una nueva entrada a la dimensión de provincias, para la provincia de A Coruña, debemos modificar el último registro de A Coruña (éste tiene como 'fecha_fin' 2999/01/01), cambiando el valor de su campo 'fecha_fin' por un día antes del valor que tenga el campo 'fecha_inicio' del registro nuevo de A Coruña.
- Se debe indicar en el CSV la 'fecha_inicio' y 'fecha_fin' del nuevo registro.
- La fecha_inicio del nuevo registro debe ser mayor que la fecha_inicio del antiguo registro.
- El nuevo registro debe tener 2999/01/01 como valor del campo 'fecha_fin'.

De esta forma se puede llevar un seguimiento de los datos de cada provincia a lo largo de la historia.

Con respecto al proceso ETL como tal, es similar a los dos últimos casos pero, a mayores se hace la comprobación de que la 'fecha_inicio' sea menor que la 'fecha_fin'. Si esto no se cumple se mostrará un error indicando el problema y se detendrá el proceso ETL y no se realizará ningún cambio en la dimensión.

Ambas tablas, tanto la dim_provincia como en la dim_area_sanitaria, tienen un proceso ETL idéntico en cuanto a los elementos que se utilizan, los pasos que realizan y sus comprobaciones. Ésto puede comprobarse en la Fig. 5.5 y Fig. 5.6.

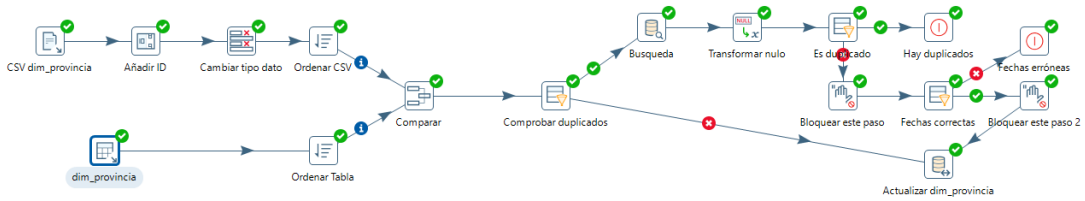

Figura 5.5: Proceso ETL Dimensión Provincia

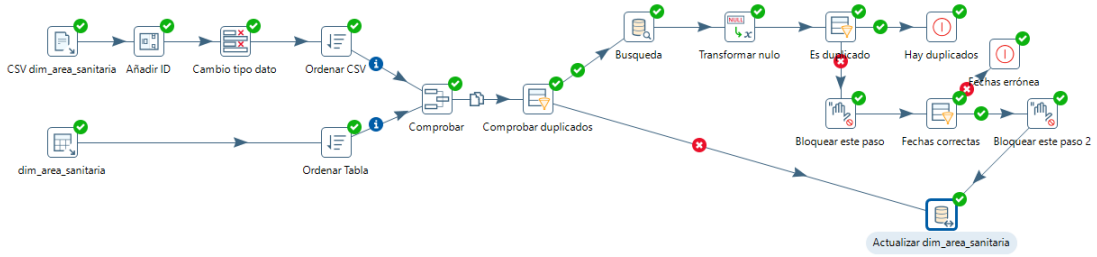

Figura 5.6: Proceso ETL Dimensión Área Sanitaria

5.5 AREA_FECHA_RESTRICCION, CASOS_GENERALES y PROV_CASOS

Los procesos ETL de las tres tablas de hechos se pueden explicar de forma conjunta ya que siguen los mismos pasos.

En primer lugar, como en cualquier proceso ETL se comienza leyendo el dato de origen, en este caso un archivo CSV. En el caso de la tabla prov_casos es necesario realizar un cambio de formato del campo fecha Fig. 5.7. En las otras dos tablas de hechos, los datos no necesitan pasar por un proceso de tranformación.

Posteriormente se comparan los registros del CSV con los que existen en la tabla de hechos del Data Warehouse y, dependiendo del resultado de la comparación se marca el registro, al igual que se comentó en las dimensiones, como identical, deleted, changed (pensado para errores en inserciones) o new (este valor se guarda en un campo nuevo que no se carga en el Data Warehouse).

Por último, se actualizan los datos de cada tabla con las modificaciones, inserciones y/o borrados correspondientes.

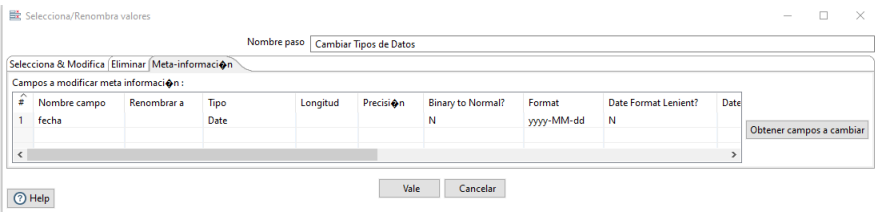

Figura 5.7: Paso del proceso ETL usado para cambiar el formato de la fecha a yyyy-MM-dd, el estándar del Data Warehouse

Cabe recalcar que, en este caso, la clave primaria de las tablas está formada por:

- fecha-id_area_sanitaria-restriccion en la tabla Area_Fecha_Restriccion (Fig. 5.8)
- fecha-id_provincia para el caso de Prov_Casos (Fig. 5.9)
- fecha-id_area_sanitaria en Casos_Generales (Fig. 5.10)

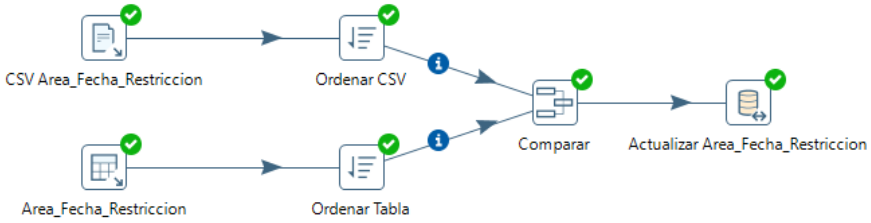

Figura 5.8: Proceso ETL tabla de hechos Area_Fecha_Restriccion

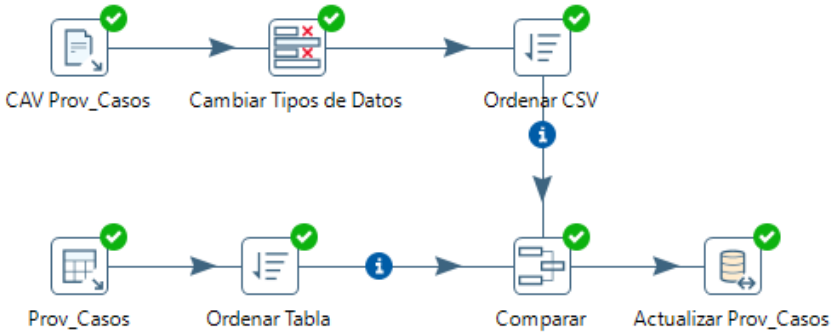

Figura 5.9: Proceso ETL tabla de hechos Prov_Casos

Es por ellos por lo que no es necesario generar un identificador para cada registro desde la herramienta y, por ende, el proceso ETL no tiene que ser el encargado de realizar una comprobación de si se está intentando insertar un registro que hace referencia a uno ya existente,

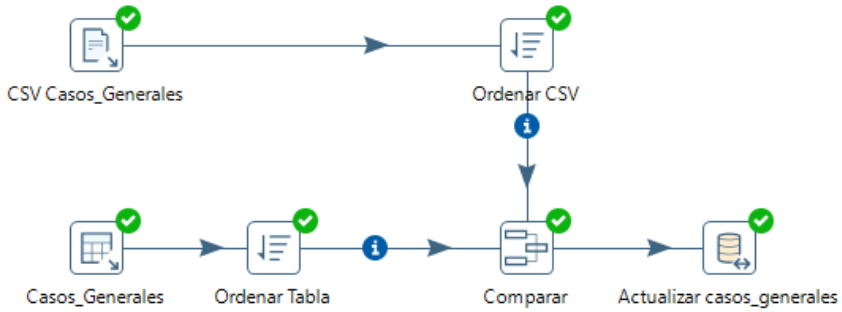

Figura 5.10: Proceso ETL tabla de hechos Casos_Generales

es decir con la misma clave primaria. En este caso es la propia tabla la que no permite la inserción debido a que se está intentando añadir un registro con una clave primaria ya existente y su inserción violaría la restricción de clave primaria de la tabla de hechos.

Capítulo 6

Explotación del Data Warehouse

Una vez generado el Data Warehouse, comienza la etapa de creación de informes que permitirán sacar conclusiones sobre los datos analizados.

Sabiendo los objetivos fijados para el estudio y los datos que conforman el Data Warehouse, se deciden los informes que se realizarán en base a dichos objetivos y la utilidad que tendrían cada uno.

6.1 Estudio Casos Provincias España

A pesar que el estudio se basa en el impacto de la pandemia en territorio gallego, se pueden aprovechar los datos de las demás provincias de España para poder tener una visión general de cómo ha afectado el COVID-19 en el país y compararlo con la situación en Galicia. De esta forma, con el siguiente informe se busca conocer en qué provincias de España han habido más casos positivos (Fig. 6.1) y, de esos casos, cuántas personas han fallecido (Fig. 6.2).

Con estos informes se puede ver claramente que hay dos provincias que superan de manera notoria a las demás en cuanto a número de contagios y defunciones, Madrid y Barcelona. Es algo normal ya que son las dos provincias de España con más habitantes con diferencia y, por ello, en el caso de que todas las provincias estuviesen igual de afectadas, son los lugares donde existirían más contagios y defunciones. En la Fig. 6.3 se puede apreciar la diferencia de población de Madrid y Barcelona con respecto a las demás provincias.

Como este estudio se centra en Galicia, se ha querido comparar la provincia gallega con mayor número de casos, muertes, hospitalizaciones e ingresos en UCI con otras provincias españolas que tengan poblaciones similares. El objetivo es ver cómo se ha controlado la pandemia en la provincia, a priori, más problemática del territorio gallego, en comparación con otras de diferentes zonas del país con poblaciones similares.

Para realizar esta comparación de una forma sencilla, se elabora un nuevo informe (Fig. 6.4) en el que se muestran los números de contagios, muertes, hospitalizaciones e ingresos en UCI

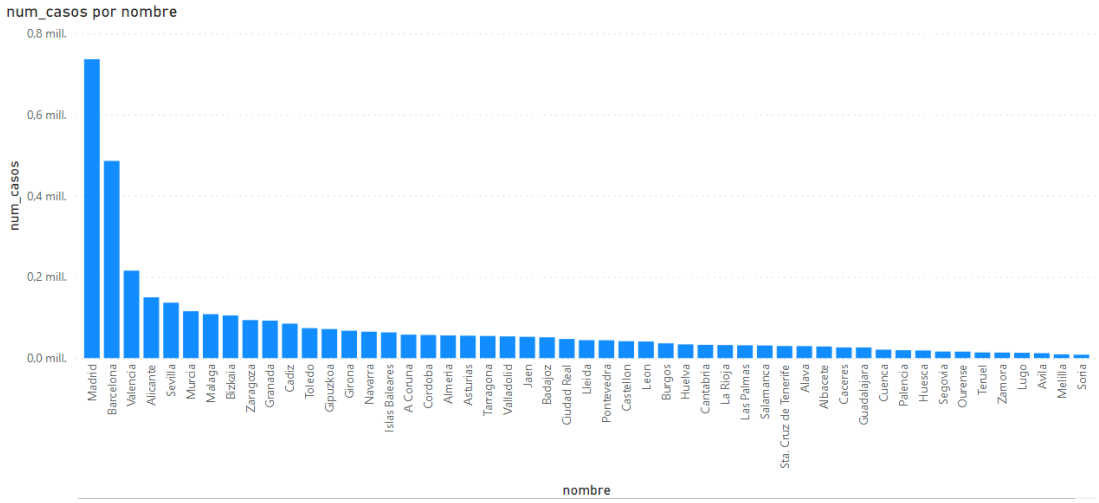

Figura 6.1: Informe que representa el número de casos positivos de cada provincia de España por COVID-19 hasta el 05/07/2021

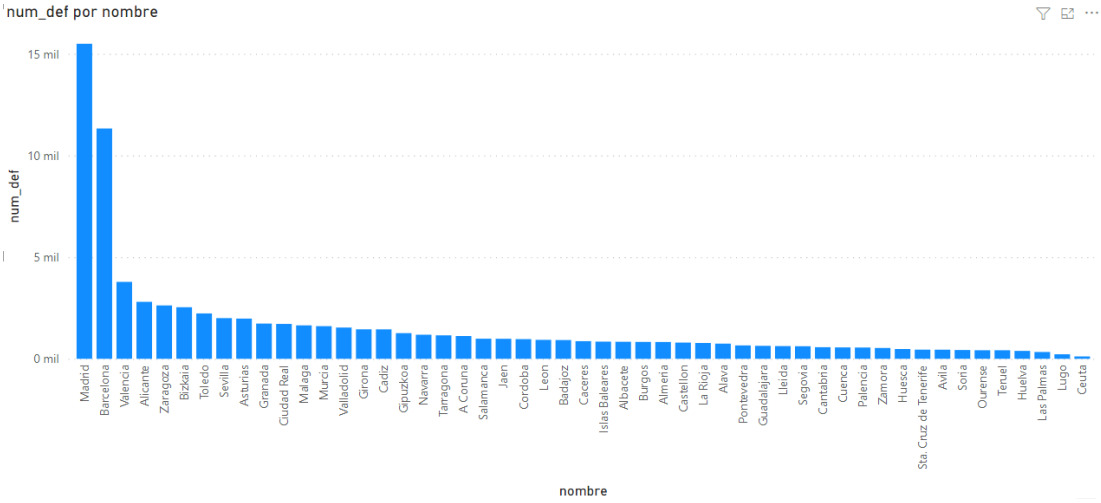

Figura 6.2: Informe que representa el número de muertes de cada provincia de España por COVID-19 hasta el 05/07/2021

para las provincias con una población mayor de 1.000.000 y menor de 1.500.000 habitantes.

A pesar que con el informe de la Fig. 6.4 se pueden apreciar ciertas diferencias entre las provincias, es probable que si se realizan las mismas comparaciones teniendo en cuenta la población, los resultados sean distintos. Esto se debe a que, a pesar de que tienen poblaciones similares, no son exactamente iguales. Con el siguiente informe, disponible en la Fig. 6.5, se busca realizar una comparativa más exacta, teniendo en cuenta, como se ha dicho, el número de habitantes de cada provincia.

Con este informe se determina que la provincia gallega no ha sobresalido ni positiva ni

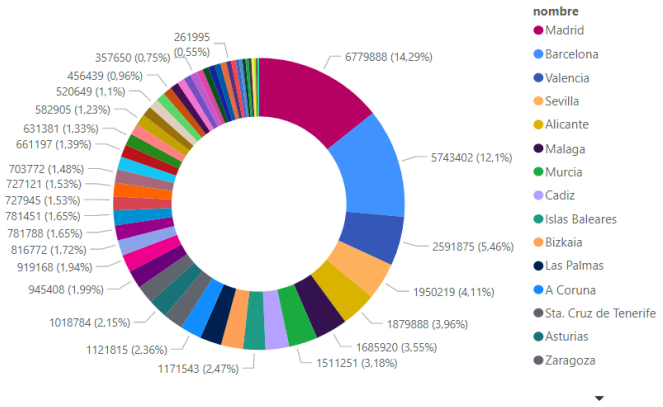

Figura 6.3: Población de las provincias españolas y el porcentaje de habitantes de España que residen en cada una

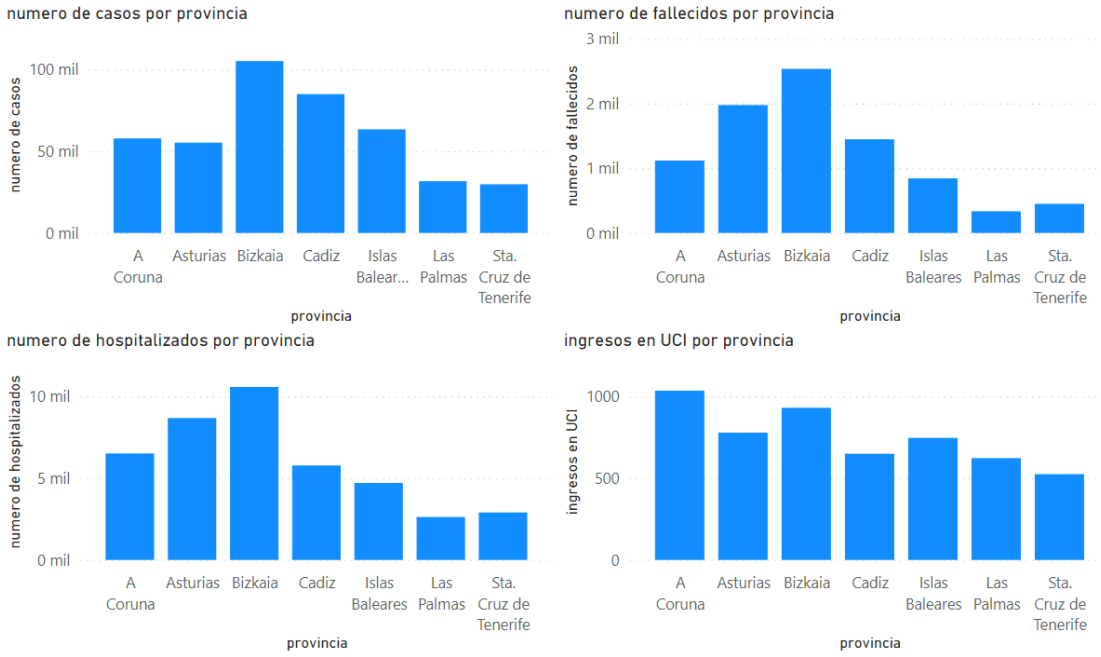

Figura 6.4: Informes comparativos de provincias con número de habitantes similares

negativamente en el control de la pandemia ya que sus cifras están rondando la media en números de contagio y muertes en comparación con las demás provincias del informe. El tema de hospitalizaciones puede deberse a la alta natalidad de la población.

6.2 Datos generales provincias Galicia

A la hora de comprender el impacto que ha tenido el COVID-19 en el territorio gallego, se debe realizar un informe que muestre los datos generales que ha producido la pandemia

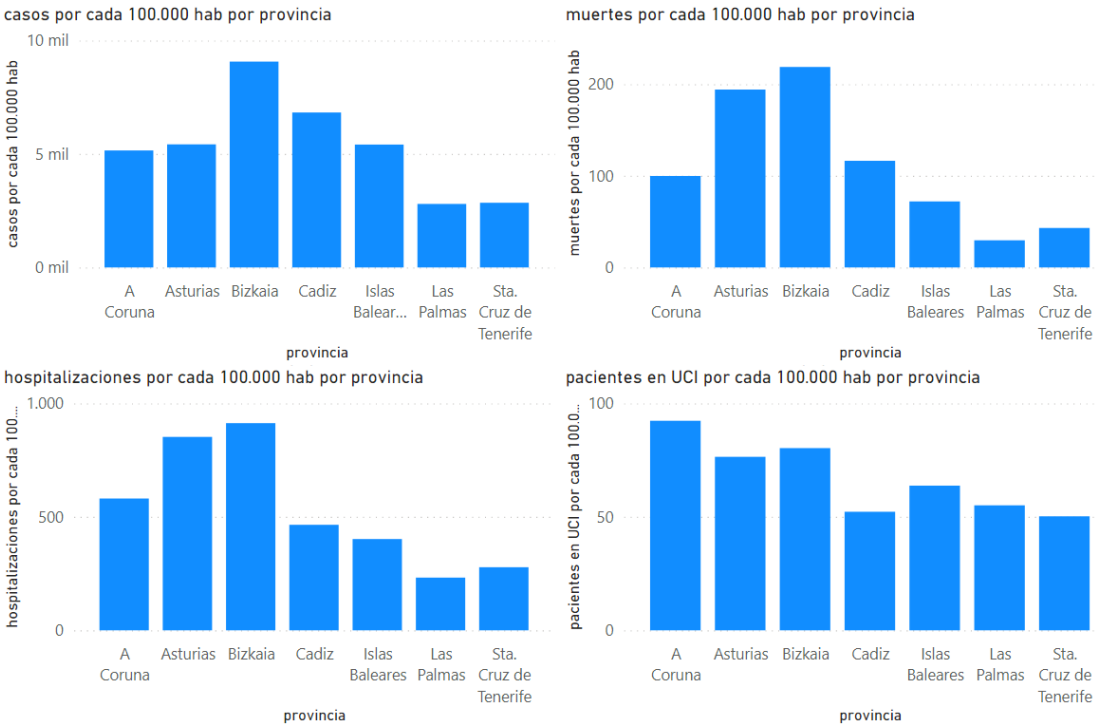

Figura 6.5: Informes comparativos de provincias con número de habitantes similares (con datos por cada 100.000 habitantes)

en Galicia (Fig. 6.6). De esta forma se busca ver cuáles han sido las provincias más y menos afectadas teniendo en cuenta el número de contagios, defunciones y personas hospitalizadas o incluso en el área de cuidados intensivos.

Se puede apreciar también la población de estas provincias que, siguiendo el ejemplo anterior, invita a pensar (sin tener conocimiento previo) que las provincias de A Coruña y Pontevedra, serán las que tengan mayor número de habitantes debido a que sus números (contagios, muertes, etc.) durante la pandemia son más altos que las otras dos provincias a analizar. De esta forma, gracias al informe de la Fig. 6.7 se puede apreciar la diferencia de población de las provincias pertenecientes a Galicia.

El papel que desempeña la población (número de habitantes) es muy interesante ya que, que una provincia tenga más casos, no quiere decir que fuese la que más cantidad de casos acumularía si todas tuviesen la misma población. Esta es una suposición interesante y, para apoyarla, se genera un informe que muestra, por cada 100.000 habitantes, el número de casos (o contagios), fallecidos, hospitalizados e ingresos en UCI, todo ello dividido por provincias, como se puede apreciar en la Fig. 6.8.

Gracias al informe de la Fig. 6.8, que muestra los datos por cada 100.000 habitantes, se puede ver con claridad cómo cambian los datos con respecto al informe de la Fig. 6.6. Al no

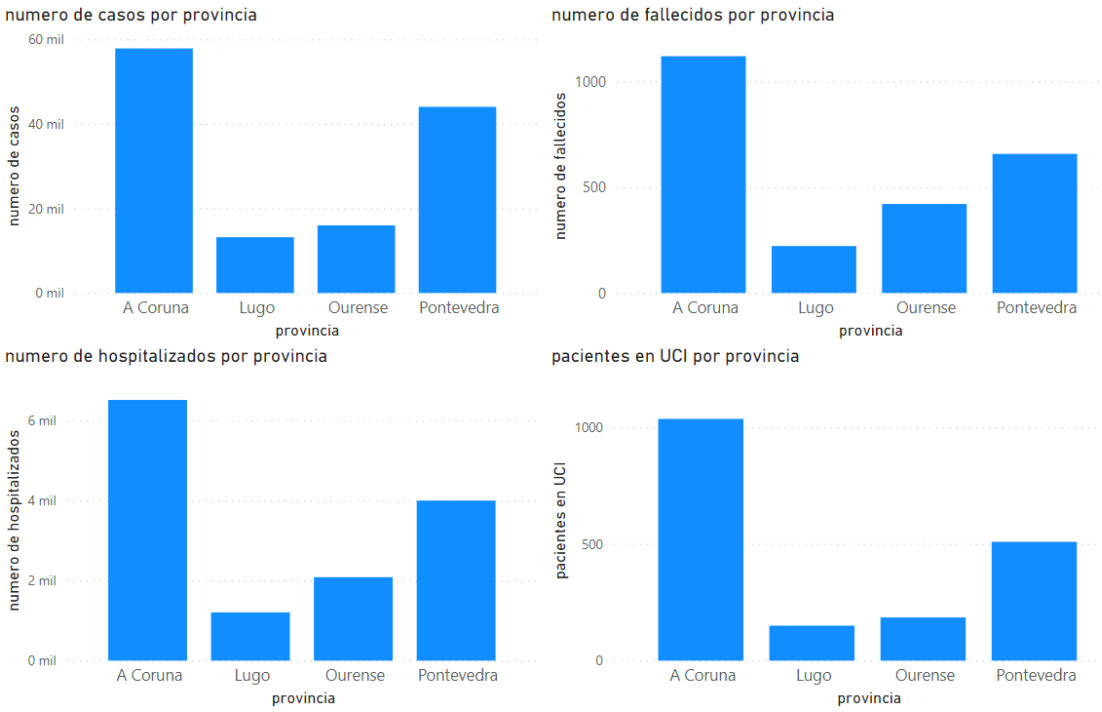

Figura 6.6: Números totales de casos positivos, fallecidos, personas hospitalizadas y personas ingresadas en UCI hasta 05/07/2021 en cada provincia de Galicia

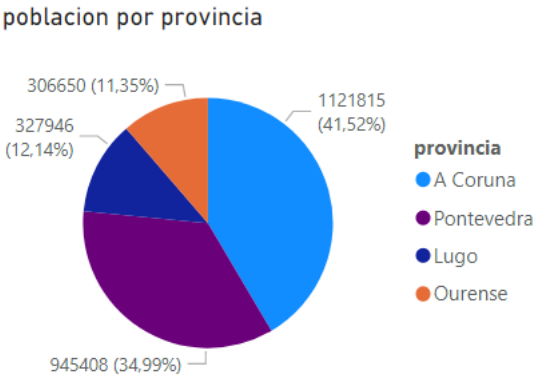

Figura 6.7: Número de habitantes por provincia en Galicia

tener en cuenta el número de habitantes, las provincias con mayor población (A Coruña y Pontevedra) acumulan una cantidad de contagios, fallecimientos, hospitalizaciones e ingresos en UCI, muy superior a la de las provincias con menor población (Ourense y Lugo). Sin embargo, si se tienen en cuenta los habitantes de cada provincia y se realizan los cálculos en base a la población (en este caso por cada 100.000 habitantes), se puede apreciar que los números cambian drásticamente y, aunque todas tienen más o menos los mismos resultados, la provincia a la que más le afecta el virus (por habitante) es, casualmente, una de las que menor

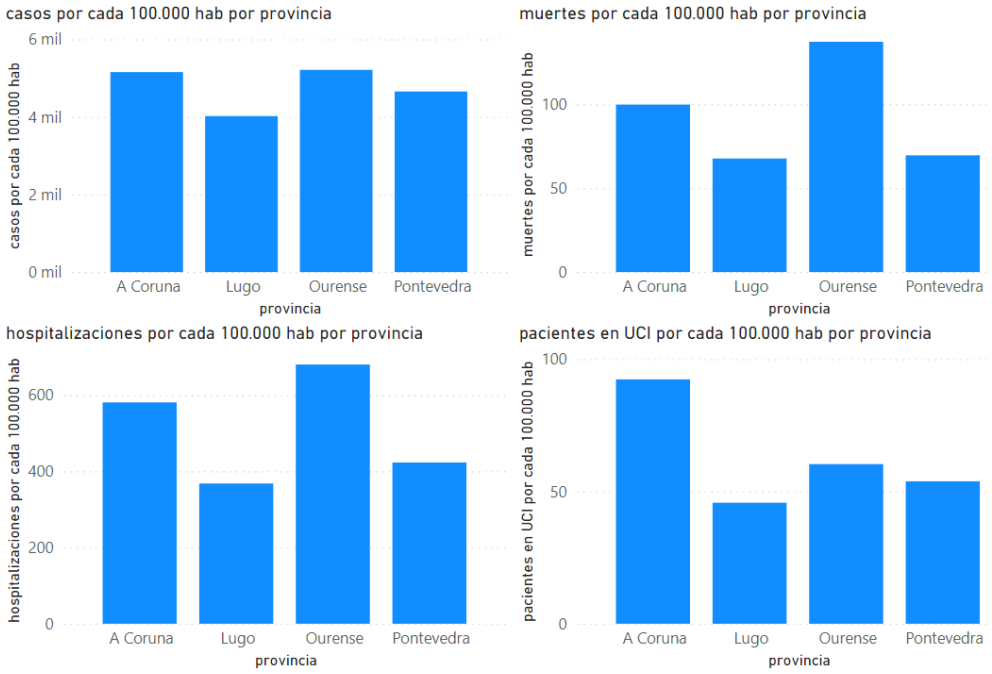

Figura 6.8: Datos totales de casos positivos, fallecidos, hospitalizaciones y personas ingresadas en UCI por cada 100.000 habitantes (hasta el 05/07/2021)

población tiene (Ourense). A pesar de que Lugo sigue ocupando la última plaza en cuanto al impacto de los contagios, los resultados comparándolo con las provincias son mucho más parejos que si no se tuviese en cuenta la población.

Con unos números tan similares, se puede concluir que el virus ha afectado a las provincias de gallegas de una forma bastante parecida, a excepción de Lugo, que tiene unos números ligeramente más bajos que el resto.

Además, viendo los resultados de los anteriores informes, se puede determinar que los lugares en los que más contagios ha habido, son los mismos en los que ha habido más defunciones, hospitalizados y también ingresos en UCI. Es decir, teniendo en cuenta el conjunto de datos con el que se opera, el número fallecidos, de pacientes hospitalizados y de pacientes en UCI son directamente proporcionales al número de contagios (o casos).

6.3 La importancia del Sexo

Continuando con el análisis de Galicia, se analiza si realmente el sexo puede llegar a ser un factor determinante que permita saber si es más o menos probable que fallezca una persona.

Se han hecho varios estudios a lo largo de la pandemia y, según éstos, el sexo sí es un factor diferencial ya que hablan de que los hombres tienen más posibilidades de fallecer que las mujeres al contraer el virus.

De esta forma se realiza un análisis que permite comprobar si los datos del Data Warehouse pueden respaldar esta suposición para las provincias de Galicia (Fig. 6.9).

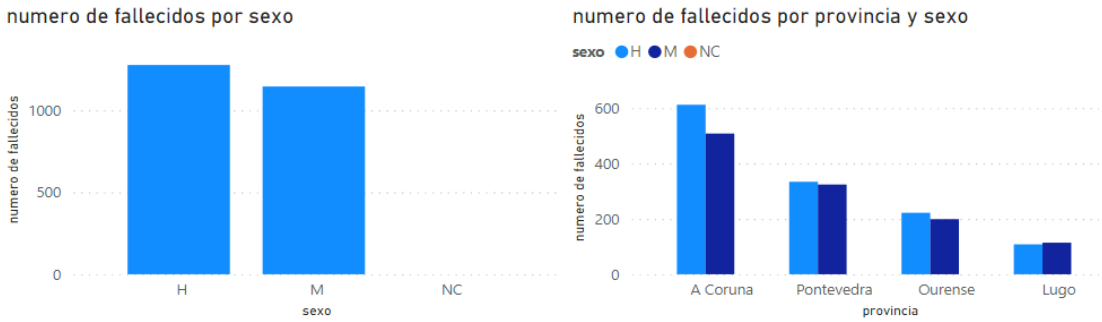

Figura 6.9: Comparativa de número de fallecidos por sexo (a nivel Comunidad Autónoma gráfico izquierdo, a nivel provincia gráfico derecho)

En ambos gráficos se aprecia que la diferencia entre los dos sexos no es muy grande pero sí existente. En el caso de Lugo, fallecen más mujeres que hombres pero, se puede decir que, por norma general, hay más fallecidos del sexo masculino que femenino.

Con esto se puede determinar que mueren más hombres que mujeres pero no se puede concluir que la probabilidad de fallecimiento a causa del virus sea mayor para los hombres que para las mujeres ya que es posible que haya muchos más casos positivos en COVID-19 del sexo masculino. Por esta razón se realiza un análisis que permita conocer qué sexo es el que engloba más casos positivos (Fig. 6.10). Con él se puede ver claramente cómo el sexo con más casos es el de las mujeres y, por ello se puede concluir que el COVID-19 sí parece ser más mortal para hombres ya que a pesar de tener menos casos positivos, el número de fallecidos es mayor.

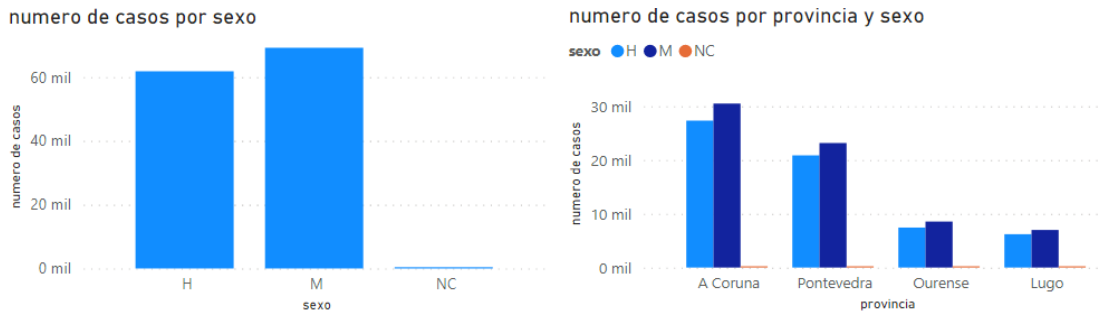

Figura 6.10: Comparativa de número de casos positivos en COVID-19 por sexo (a nivel Comunidad Autónoma gráfico izquierdo, a nivel provincia gráfico derecho)

Gracias a estos informes que se acaban de realizar se puede apreciar que el sexo masculino es el que menos se contagia y el que más fallecimientos acumula en Galicia, pero sería interesante poder obtener la probabilidad (o porcentaje de posibilidades) que tiene un paciente de

fallecer al contagiarse de COVID-19 en las diferentes provincias. Para ello se crea un informe que muestre dicha probabilidad en base al sexo y localización (provincia) del individuo (Fig. 6.11).

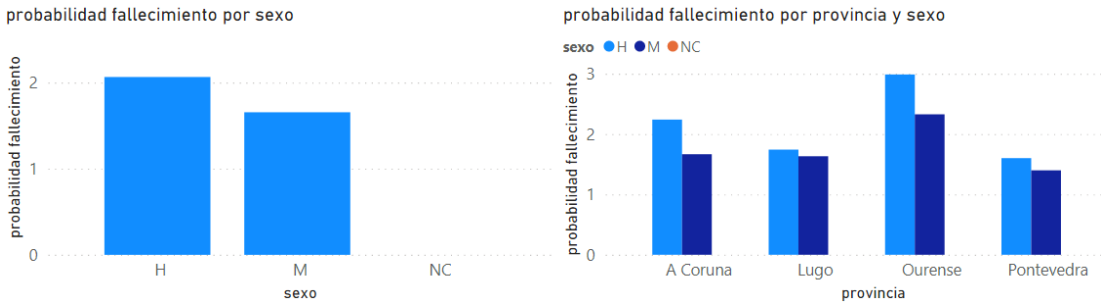

Figura 6.11: Comparativa de la mortalidad para cada sexo en tanto por ciento (a nivel Comunidad Autónoma gráfico izquierdo, a nivel provincia gráfico derecho)

Con este último informe se entiende mejor la importancia del papel que juega la población y que, a pesar de que Ourense no es la provincia con más fallecidos, sí es la que posee el mayor porcentaje de mortalidad tanto para hombres como para mujeres (menos del 2% en ambos casos).

6.4 La importancia de la edad

Además del sexo de una persona, también es interesante conocer lo importante que puede llegar a ser el factor edad en cuanto a la mortalidad. Al igual que se pudo concluir qué sexo era el más afectado desde el punto de vista de los fallecimientos, se realiza el estudio equivalente para los grupos de edades (Fig. 6.12).

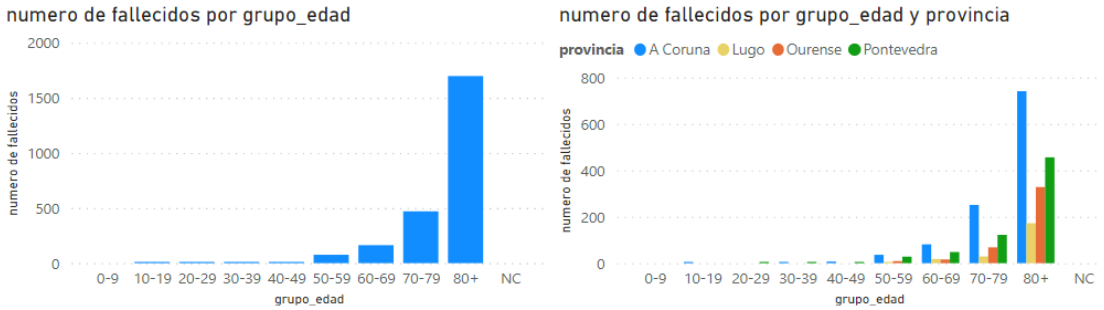

Figura 6.12: Comparativa de número de defunciones por COVID-19 agrupados por grupos de edades (a nivel Comunidad Autónoma gráfico izquierdo, a nivel provincia gráfico derecho)

En este caso se puede apreciar, con mucha más claridad que en el análisis anterior (importancia del sexo del paciente), que el grupo de edad más afectado (el que más muertes acumula) es el correspondiente a las personas de más de 80 años. A mayores, cabe resaltar la existencia

de una clara curva ascendente a medida que se avanza por los rangos de edades (de menor a mayor edad). Por consiguiente, parece que a las personas mayores les afecta mucho más el virus que a las personas de mediana edad, jóvenes y/o niños.

Al igual que en el caso anterior, también es conveniente revisar si el gran número de fallecidos que representan los grupos de edades más avanzadas se puede corresponder con un alto número de casos positivos en COVID-19. Por ello, se genera un nuevo informe (Fig. 6.13) con el fin de tener una visión general de las edades que han acumulado más contagios y , además, poder utilizarlo para determinar si el elevado número de fallecidos de un grupo de edad, puede deberse a un gran número de contagios del mismo.

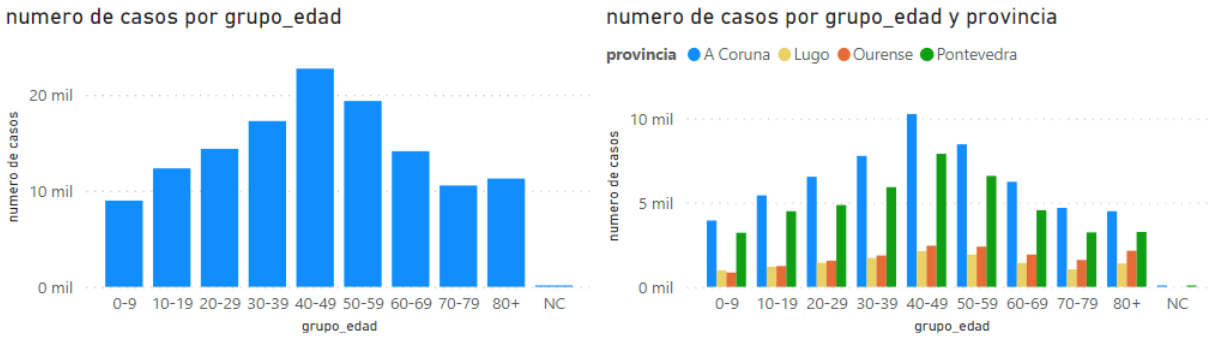

Figura 6.13: Comparativa de número casos positivos en COVID-19 agrupados por grupos de edades (a nivel Comunidad Autónoma gráfico izquierdo, a nivel provincia gráfico derecho)

Al analizar los datos que presentan estos dos informes, se puede ver que, a pesar de sus altísimos números de personas fallecidas, los grupos de edad más avanzados son de los que menos contagios acumulan. También cabe recalcar que los grupos de edades pertenecientes a personas de mediana edad, desde los 20 años hasta los 50/60 son los que han acumulado más contagios mientras que son los extremos (los más mayores como los más jóvenes) los que representan menos casos.

Además de todo esto, yendo un poco más allá en el análisis, sería muy interesante saber cuál es la provincia que tiene la probabilidad más alta de mortalidad para una persona que ha sido contagiada (basándose en el conjunto de datos usado en el estudio). Para profundizar en este tema, se realiza un informe que permita visualizar el porcentaje de personas contagiadas que han fallecido haciendo una distinción por provincias y grupos de edades (Fig. 6.14), permitiendo ver en qué provincia fue más letal el virus y también el porcentaje de mortalidad del virus para cada grupo de edad.

En este nuevo informe (Fig. 6.14) se puede apreciar con claridad que el grupo de edad con mayor probabilidad de muerte al contraer el COVID-19 es el de mayores de 80 años, superando en todas las provincias el 10% de probabilidad de fallecimiento y, en dos de ellos, incluso llegando a sobrepasar el 15%.

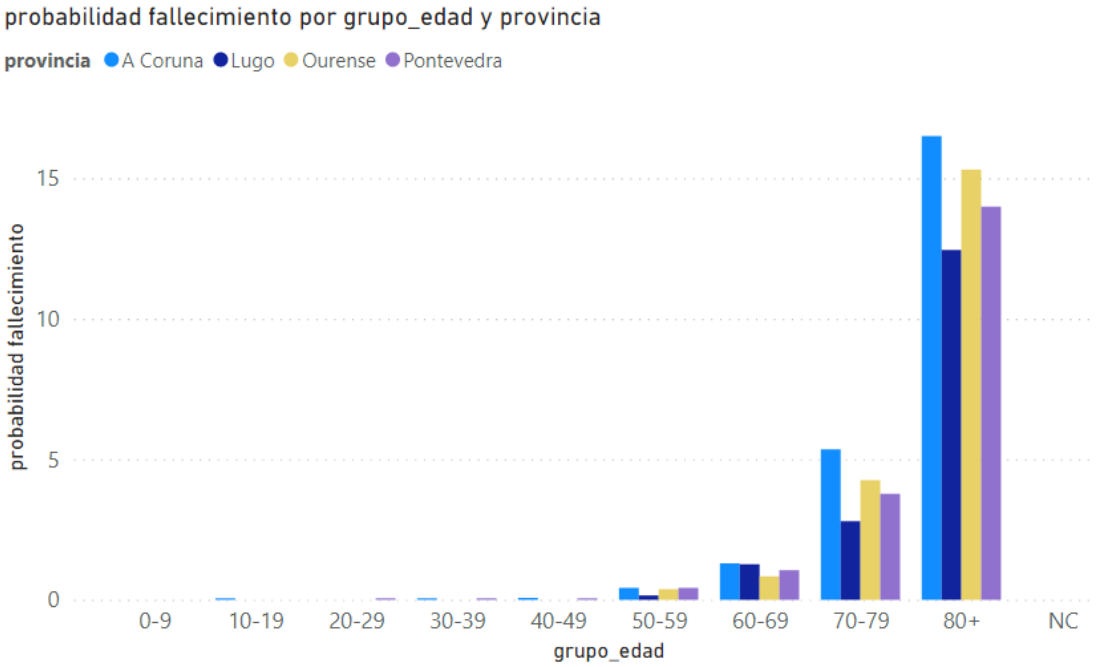

Figura 6.14: Comparativa de la mortalidad en los diferentes grupos de edad para cada provincia

La diferencia entre este grupo de edad y el anterior con más probabilidad (70-79) es excesivamente grande, llegando a triplicar el porcentaje de este último. Cabe recalcar que, gracias a este gráfico se puede apreciar mucho mejor el impacto del virus en los diferentes rangos de edades, tanto el que tiene en los mayores de 80 años (con preocupantes probabilidades de fallecer), como el que tiene en los menores de 70 años, cuya probabilidad de fallecer es mucho menor (prácticamente probabilidad nula, por debajo de los 50 años).

Gracias a combinar la información obtenida de los informes que representan el número de fallecidos y los que muestran el número de contagios, se puede ver que la edad sí es un punto a tener muy en cuenta a la hora de saber si una persona tiene más o menos posibilidades de fallecer por culpa del virus.

6.5 Hospitalizacion de pacientes

Al tener datos sobre la cantidad de personas que estuvieron hospitalizadas y las que estuvieron en la unidad de cuidados intensivos (UCI), es interesante utilizar dichos datos para saber cuales fueron las personas que necesitaron más atención médica ordinaria y, sobre todo, si la unidad de cuidados intensivos tuvo algún tipo de paciente en concreto (Fig. 6.15 y Fig. 6.16). Es decir, saber (independientemente de si fallecen o no) a qué grupos personas les afectaron más los síntomas hasta ahora.

Con ellos se puede ver que la mayoría de personas ingresadas en el hospital son aquellas

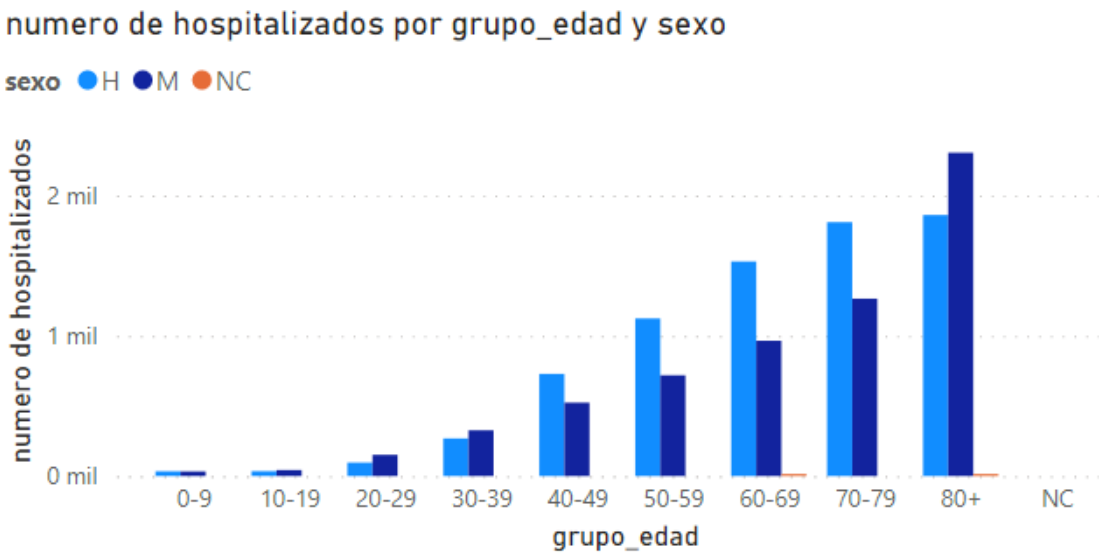

Figura 6.15: Comparativa de número de hospitalizados por COVID-19 agrupados por grupos de edades

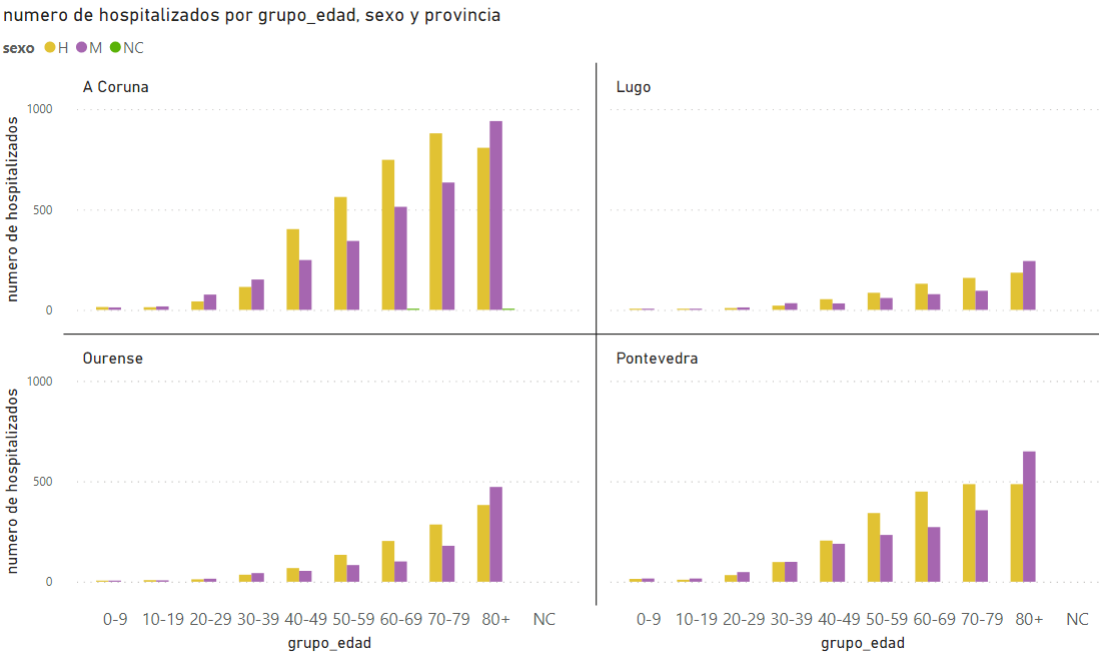

Figura 6.16: Comparativa de número de hospitalizados por COVID-19 agrupados por grupos de edades y provincia

que tienen mayor tasa de mortalidad, es decir, la gente mayor. De hecho, el orden grupos de edades más hospitalizados coincide con el orden de los grupos de edades con mayor cantidad de muertes. Además, en cuanto al sexo se refiere, se puede distinguir una predominancia de personas del sexo masculino en la mayoría de grupos de edades a excepción del grupo que

recoge a las personas más mayores, los que superan los 80 años, donde las mujeres son las que más ingresan en los hospitales.

Sabiendo que la gente que ingresa en el hospital es aquella que está más grave de lo normal, los resultados de los anteriores informes, son lógicos ya que las personas mayores son las más afectadas por el virus y, por tanto, las que más lo sufren, por ello son las más ingresadas.

6.6 Unidad de Cuidados Intensivo (UCI)

Por otro lado, la unidad de cuidados intensivos es un área del hospital con plazas limitadas y que no puede atender a todos los pacientes. En ocasiones, deben tomar decisiones de si llevar, o no, a una persona a dicha área dependiendo de sus posibilidades de sobrevivir y de cómo ésta se puede recuperarse tras el tratamiento.

Un buen comienzo para este análisis sería ver cual es la probabilidad general de ingresar en la UCI una vez está hospitalizado un paciente, para ello se genera un nuevo informe (Fig. 6.17) en el que se aprecia que una pequeña parte de los hospitalizados por COVID-19 son ingresados en el área de cuidados intensivos, por lo que, lo más normal es que un hospitalizado sea dado de baja sin haber necesitado ingresar en la UCI.

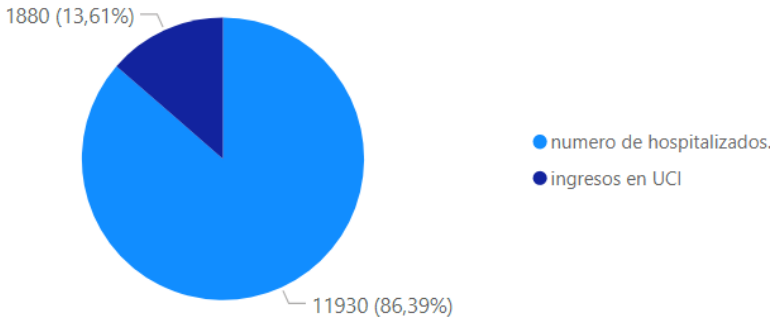

Figura 6.17: Porcentaje ingresos en UCI tras haber sido hospitalizados por COVID-19

Profundizando un poco más en el análisis de ingresos en UCI, los dos siguientes informes (Fig. 6.18) permiten contemplar los números de los grupos de edades que han sido atendidos en la UCI.

Es sencillo darse cuenta de que, a pesar de que las personas que superan los 80 años son las que más ingresan en el hospital, no es el mismo caso en la unidad de cuidados intensivos. Como es gente mayor, muchas veces no llegan a dicho área, los médicos deciden no llevarlos por sus protocolos, etc.

Desde el punto de vista de la edad, se podría pensar que al ser los que tienen más mortalidad y los que más ingresan en hospitales, debería ser también el grupo de edad que más se daría de alta en la UCI. Pero, por el contrario, es un grupo de edad que no se caracteriza por

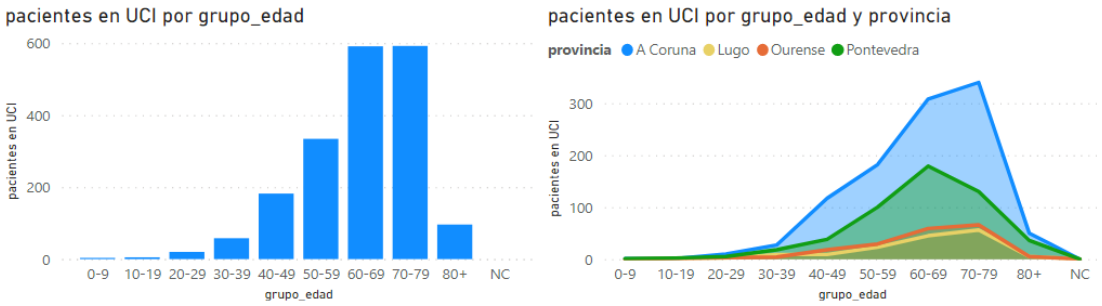

Figura 6.18: Comparativa de número de ingresados en UCI por COVID-19 agrupando por grupos de edades

las altas en este área.

Sin embargo, se ve que las personas mayores (pero menores de 80 años), sí son las que más ingresan en la UCI, esto podría ser debido a que siguen siendo grupos de gente mayor y con una tasa de mortalidad alta, por lo que pueden tener síntomas más agravados, y no son tan mayores como el otro grupo. Es posible que por ello puedan soportar mejor que las personas mayores de 80 el tratamiento más intensivo y agresivo de la UCI, por lo que las altas, son mucho mayores. Con el fin de aportar todavía más información de interés, se genera otro informe que permita saber las posibilidades que tiene una persona ya hospitalizada de ingresar en UCI teniendo en cuenta el grupo de edad al que pertenece (Fig. 6.19).

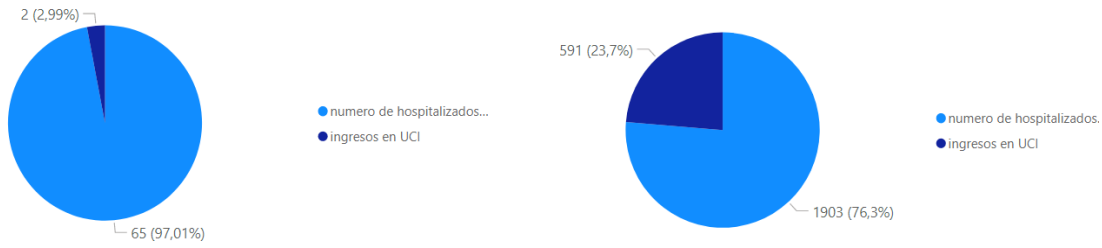

Figura 6.19: Porcentaje ingresos en UCI tras haber sido hospitalizados por COVID-19 (izquierda grupo 0-9 años, derecha grupo 60-69 años)

En el informe de la (Fig. 6.19) se han tratado dos casos, en el de la izquierda de la figura se refiere al grupo de edad 0-9 años y el de la derecha al de 60-69. Esto se ha hecho para dejar clara la gran diferencia que hay entre el grupo con menor y mayor probabilidad.

Parece entonces que los individuos que ingresan en la UCI, son aquellos pertenecientes a los grupos más afectados por el virus y, como se ha visto, coincide con las personas con mayor mortalidad, es por ello por lo que sería interesante ver qué sucede con los hombres ya que tienen mayor tasa de mortalidad que las mujeres. Se puede suponer que, siguiendo la lógica de los anteriores informes, como son más propensos a morir por COVID-19, los hombres sean los que encabecen el número de altas, pero se verá con mayor claridad en el

informe de la Fig. 6.20.

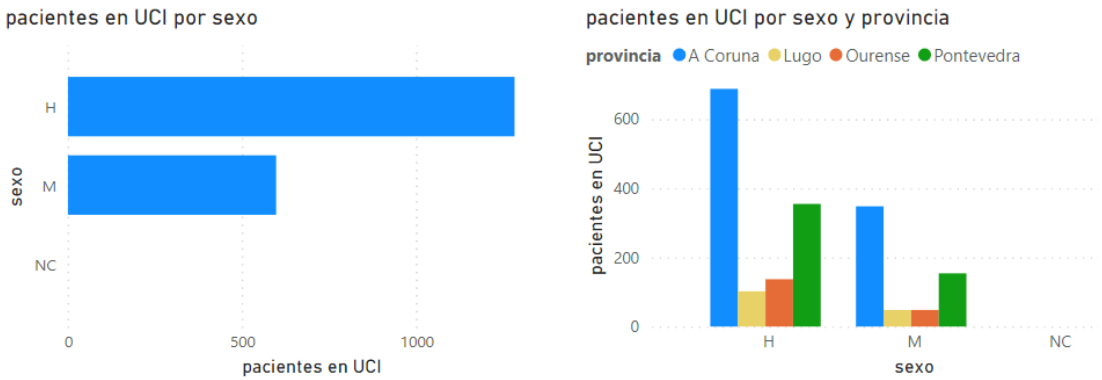

Figura 6.20: Comparativa de número de ingresados en UCI por COVID-19 agrupando por sexo

De esta forma, se puede sacar en claro que los hombres sí tienen más posibilidades que las mujeres de ingresar en la UCI. Por ello, se puede concluir que sí importa el sexo a la hora de saber si una persona tiene más o menos posibilidades de necesite ser atendido en la unidad de cuidados intensivos.

Además de esto, también sería interesante saber la probabilidad que tiene cada sexo de ingresar en cuidados intensivos una vez hospitalizado. Para ello se genera un nuevo informe (Fig. 6.21), en él se puede ver que los hombres tienen casi el doble de posibilidades que el sexo opuesto de ingresar en la UCI tras su hospitalización.

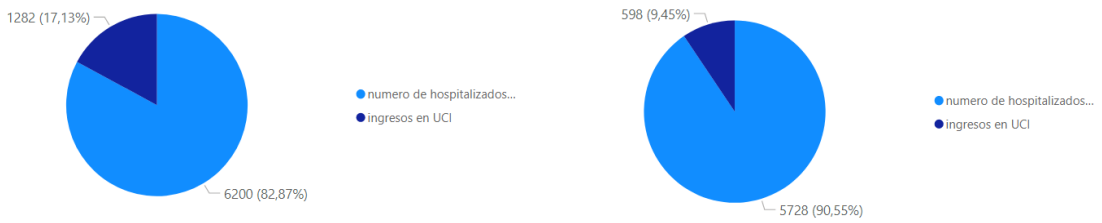

Figura 6.21: Porcentaje ingresos en UCI tras haber sido hospitalizados por COVID-19 para cada sexo (izquierda Hombres, derecha Mujeres)

De nuevo, parece que los hombres son los más afectados por el virus de entre los pacientes hospitalarios, llegando a alcanzar casi el doble de posibilidades de ingreso en la UCI por encima que las mujeres. Además, gracias a este informe también se puede determinar que la probabilidad de ingresar en la UCI una vez se ha dado de alta el paciente en el hospital es relativamente alta para los hombres (casi 20%) sabiendo la importancia que supone entrar en este área de cuidados intensivos.

6.7 Tipos de pruebas para detección de positivos

Durante la pandemia, se han realizado abundantes pruebas para poder saber si una persona era positiva en COVID-19. Esto es una realidad pero sería muy útil saber cuántas pruebas se han realizado exactamente. Para ello se hará uso de los datos recogidos. En una primera instancia se verán el número de pruebas realizadas en toda la comunidad autónoma (Fig. 6.22) agrupada por día de la semana y después se agruparán por área sanitaria (Fig. 6.23) para conocer dónde se han realizado más o menos pruebas.

dia_semana_nombre	pruebas_realizadas_pcr	pruebas_realizadas_no_pcr
Martes	267046	98066
Lunes	255933	99659
Miercoles	254073	99455
Jueves	253637	99024
Viernes	245949	74435
Sabado	225628	59345
Domingo	195727	89801
Total	**1697993**	**619785**

Figura 6.22: Informes que indican el número de pruebas PCR (izquierda) y el número de pruebas que no son PCR (derecha) hechas en Galicia (entre el 07/10/2020 y el 25/05/2021)

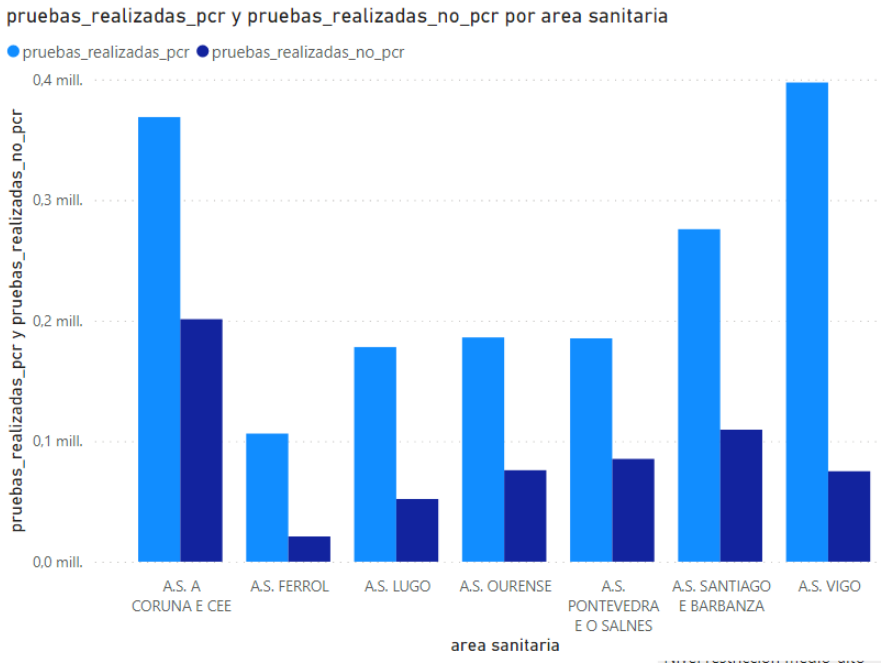

Figura 6.23: Informe sobre pruebas PCR y no PCR hechas en cada área sanitaria (entre el 07/10/2020 y el 25/05/2021)

Con estos resultados se puede divisar el gran número de pruebas realizadas en Galicia, rondando los 2,5 millones de pruebas en total. Existe una clara predominancia de las pruebas PCR sobre las demás (pruebas no PCR) en el territorio gallego y, además, los días que acumulan más pruebas son los de inicio de semana, mientras que el fin de semana es cuando se realizan menos pruebas.

Además, al realizar el informe por área sanitaria, se aprecia que, por norma general, se hacen algo más del doble de pruebas PCR que de otros tipos. Esto sucede en todas las zonas a excepción del área sanitaria de Vigo, donde el número de pruebas PCR cuadriplica los números de las demás pruebas. Es interesante remarcar que el área sanitaria de Vigo, a pesar de ser la provincia que representa el mayor número de pruebas PCR, es una de las áreas que menos pruebas no PCR acumula, por lo que parece que en esa zona de Galicia puede haber algún factor que haga que los números entre las pruebas PCR y las demás sean tan dispares.

6.8 Informes sobre el eje tiempo

Los informes realizados hasta le momento son muy interesantes ya que permiten aportar información sobre la pandemia de una forma más genérica, sin hablar de períodos o fechas concretas. Es decir, han permitido analizar la pandemia basándose en los datos totales formando algunas agrupaciones interesantes para poder deducir a qué grupos de edades, sexo, provincias, etc. afectaba más el virus, pero a partir de este punto se realizarán algunos informes que permitan conocer cómo ha sido la evolución del virus en el territorio gallego a lo largo del tiempo.

6.8.1 Contagios, Muertes, Hospitalizaciones y pacientes en UCI

A pesar de que A Coruña encabeza los gráficos con mayor número de casos positivos en COVID-19, muertes, gente hospitalizada y pacientes en UCI, estos datos no se mantienen constantes a lo largo del tiempo, existen picos y valles. Para ver cuáles han sido las etapas en donde más (y menos) se ha sufrido en cada aspecto (casos positivos, muertes, personas hospitalizadas o ingresadas en UCI) es necesario realizar una serie de informes que permitan mostrar esta información diariamente sobre cada provincia.

De esta forma se generan los siguientes informes que muestran: el número de casos positivos (Fig. 6.24), defunciones (Fig. 6.25), gente hospitalizada (Fig. 6.26) e ingresos en UCI (Fig. 6.27) (todos a nivel de provincia) desde el inicio de 2020 hasta el 05/07/2021.

El informe que se debe usar de base para analizar los otros tres es el que hace referencia a los casos positivos ya que tanto los fallecidos, como los hospitalizados y los ingresados en la UCI, son personas que son positivas en COVID-19.

Así pues, en el informe de casos positivos se puede ver bastante homogeneidad en las

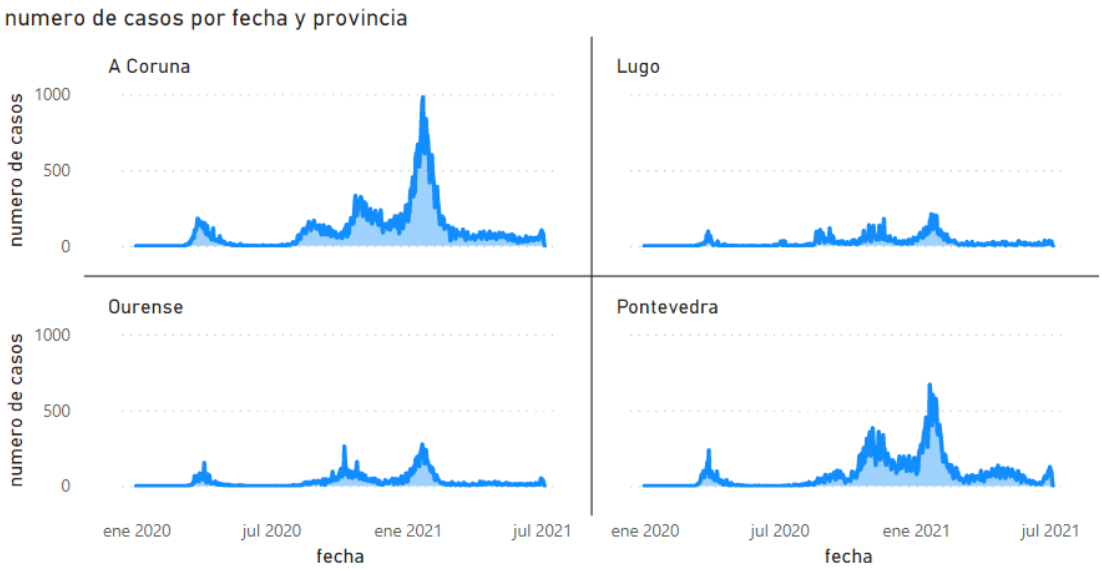

Figura 6.24: Informe sobre casos positivos en COVID-19 en cada provincia gallega a lo largo de la pandemia

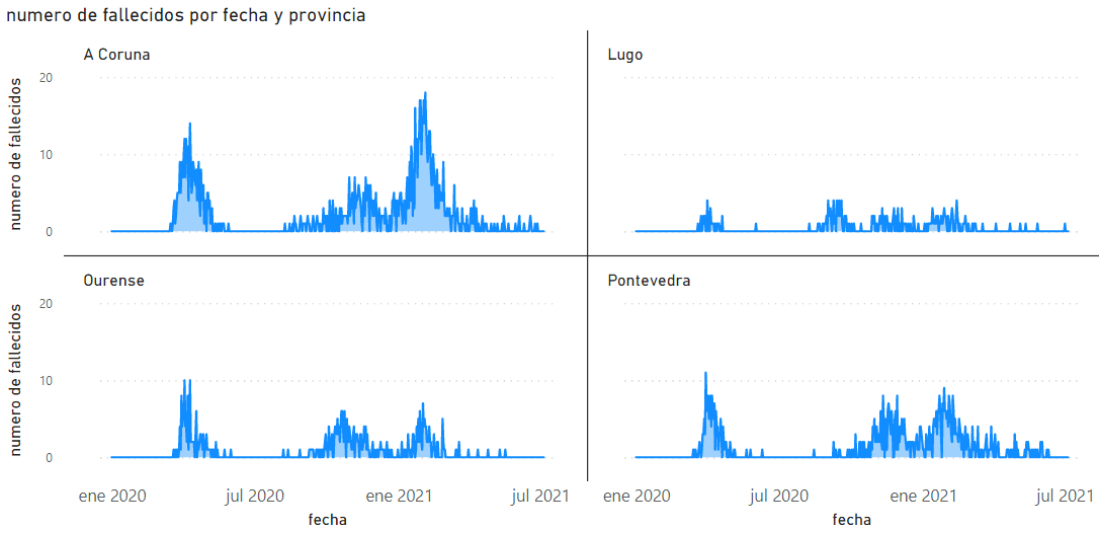

Figura 6.25: Informe sobre las muertes propiciadas por el COVID-19 en cada provincia gallega a lo largo de la pandemia

diferentes provincias en cuando a la forma de los gráficos, es decir, se aprecian subidas y bajadas en las mismas fechas, aunque en diferentes cantidades. Claramente hay dos provincias con mayor número de casos (A Coruña y Pontevedra) y otras dos que tienen muchos menos casos (Ourense y Lugo). Éste informe permite que se vean con claridad cuáles son los picos y valles, o lo que es lo mismo, en donde hubo más y menos contagios respectivamente. Claramente existe un pico más importante que el resto (sobre todo apreciable en la provincia de A

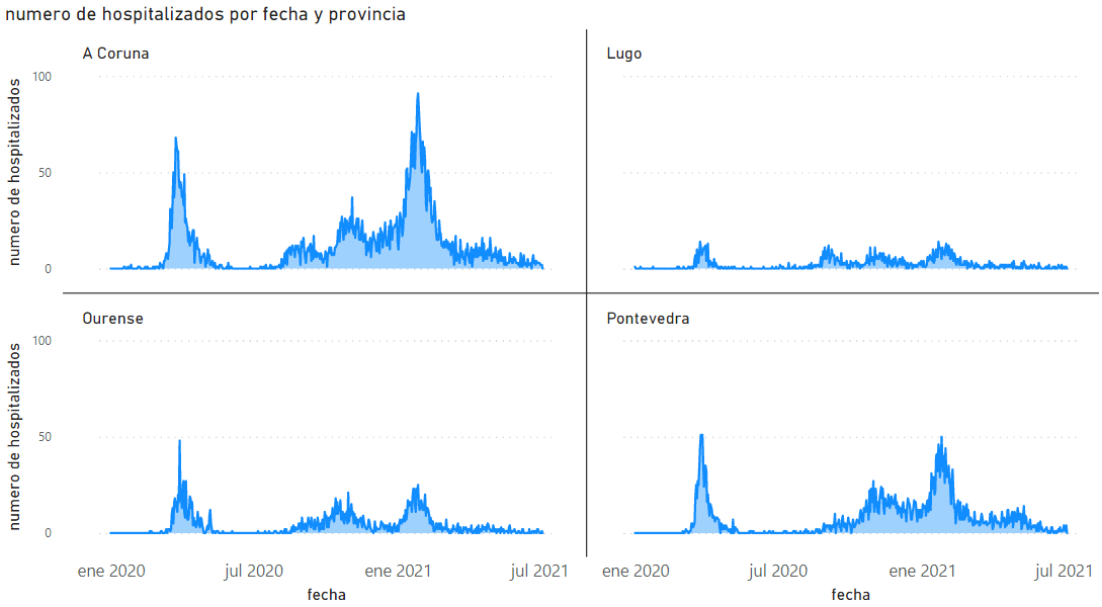

Figura 6.26: Informe sobre hospitalizaciones en cada provincia gallega a lo largo de la pandemia

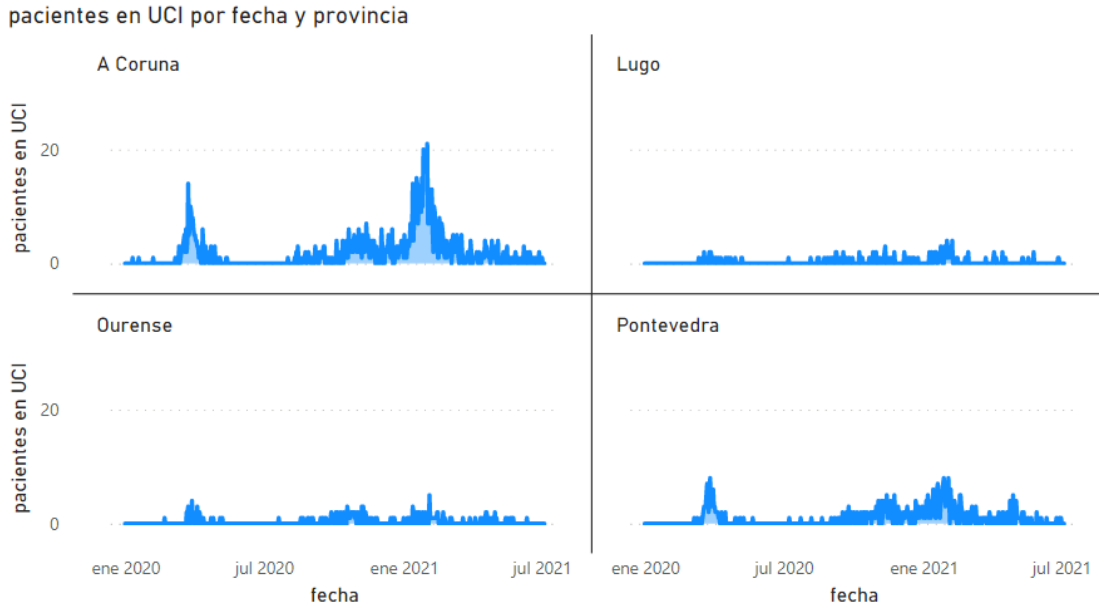

Figura 6.27: Informe sobre altas en UCI en cada provincia gallega a lo largo de la pandemia

Coruña) a finales del mes de enero y principios del mes de febrero, pero existen otros que se pueden considerar picos como el de finales de octubre/principios de noviembre (2020) y otro pico todavía más pequeño en el mes de marzo (2020).

Por el contrario, puede apreciarse que después de cada pico viene un valle en el que el

número de positivos desciende drásticamente hasta alcanzar en varias ocasiones números mínimos (tras el primer y tercer pico). Quizás el valle que más cabe recalcar es el que comprende desde mediados de abril hasta mediados de agosto, en el que realmente se consigue reducir mucho el número de contagios por día. Tras esta época de pocos casos, viene lo que será la etapa que dará pie al segundo y tercer pico (u ola), seguramente propiciada por cambiar las restricciones activas en ese momento pero esto es algo que se verá más adelante.

Sobre los demás informes, los referentes a las muertes, casos hospitalizados y personas dadas de alta en la UCI, se nota claramente la presencia de los aumentos y descensos de contagios. Evidentemente a mayor número de casos positivos (o contagios), mayor debería ser el número de personas que fallecerán e ingresarán en hospitales y unidades de cuidados intensivos.

Lo más interesante que se puede sacar de estos gráficos es que parece que los datos que se han ido recopilando desde el inicio de la pandemia, la información y experiencia asimilada han ayudado enormemente a reducir la letalidad del COVID-19 y se puede combatir el virus de una forma más efectiva que en marzo de 2020 por ejemplo.

¿Por qué podemos saber esto? Si nos fijamos en el gráfico que indica el número de casos positivos a lo largo del tiempo (Fig. 6.24), se puede apreciar cómo hay un un pequeño pico de casos sobre el mes de marzo (2020) y otro mucho mayor en enero-febrero (2021). En los otros tres informes se ven reflejados los picos de contagios, es decir, se ven claramente picos en los períodos correspondientes a los aumentos en el número de casos positivos. Pero realmente, hubo muchos más casos en enero-febrero de 2021 que en marzo del año anterior. Lo lógico sería que el número de muertes e ingresos en hospitales y UCI fuese también mucho mayor en enero-febrero que en marzo, pero esto no es así. Con ello se puede concluir que, a pesar de que los contagios han sido muchísimo mayores a inicios de 2021 que a principios de 2020, el impacto del COVID-19 en la población no ha sido tan perjudicial como al inicio de la pandemia. Los casos subieron considerablemente en enero-febrero de 2021 con respecto a la primera ola (o primer pico) pero los resultados son similares a los de marzo 2020.

6.8.2 Contagios y Muertes por Sexo

Siguiendo con la dinámica de ver algunos de los datos a lo largo del eje temporal, se continúa con un informe que representa cómo ha afectado el virus a las personas dependiendo su sexo a lo largo de la pandemia (Fig. 6.28). Es interesante saber también el número de fallecimientos diarios (Fig. 6.29) para poder ver si una ola (o pico) ha propiciado más muertes a hombres y otra a mujeres, ambas al mismo sexo, etc.

Tras ver los informes se puede concluir que no existe ningún pico en el que se hayan contagiado especialmente un sexo en concreto. Lo mismo pasa a la hora de analizar las muertes propiciadas por el virus en Galicia, en todos los casos de aumentos de fallecimientos, se

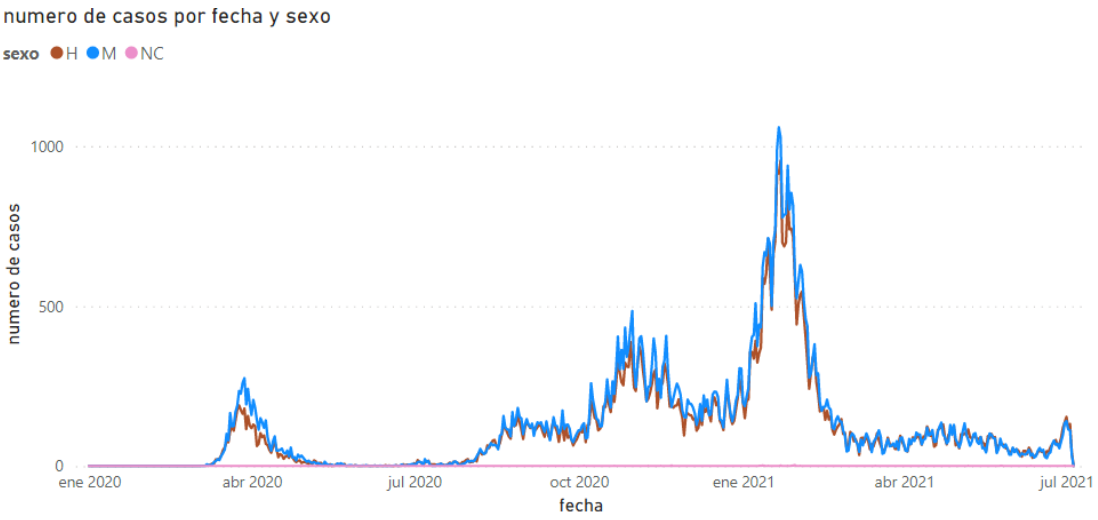

Figura 6.28: Informe de contagios diarios a lo largo de la pandemia en Galicia (a nivel sexo)

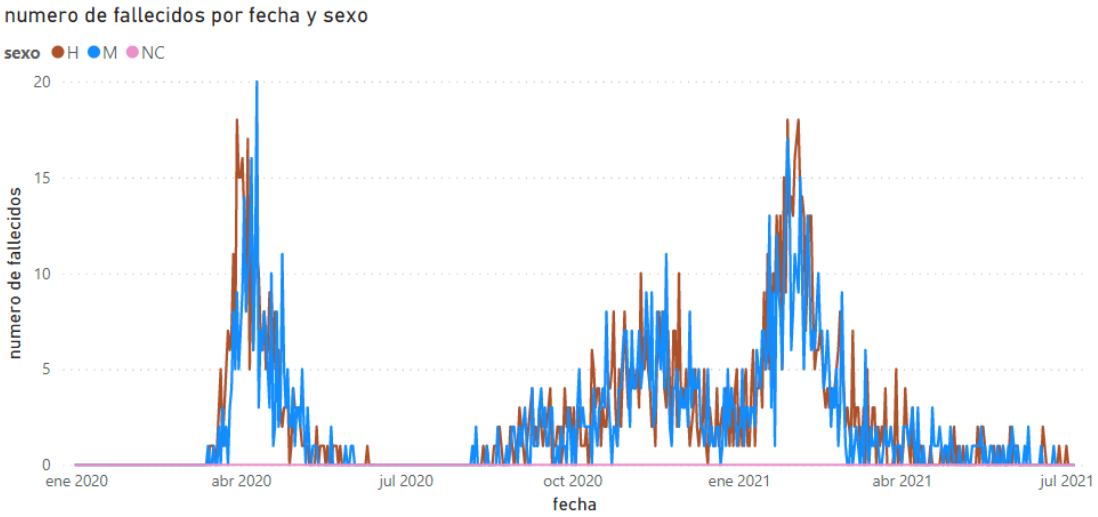

Figura 6.29: Informe de fallecimientos diarios de cada sexo a lo largo de la pandemia en Galicia

reparten los casos de hombres y mujeres de forma bastante pareja.

6.8.3 Contagios y Muertes por Edades

Siguiendo este enfoque, se procede con la generación de informes que permitan conocer el impacto que ha tenido el virus sobre los diferentes rangos de edades de forma diaria. Con esto se busca saber si ha habido períodos en los que el COVID-19 afectase más a un grupo de edad en concreto (Fig. 6.30, Fig. 6.31, Fig. 6.32, Fig. 6.33, Fig. 6.34, Fig. 6.35). También es interesante ver la reacción de grupos que no suelen estar afectados por la enfermedad ante los aumentos masivos de casos.

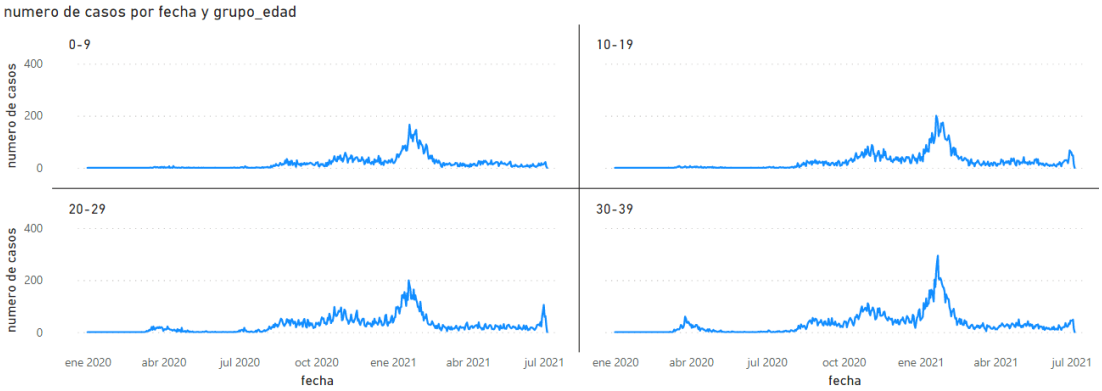

Figura 6.30: Informe que muestra, de forma diaria, el número de personas menores de 40 años que han dado positivo en COVID-19 a lo largo de la pandemia en Galicia

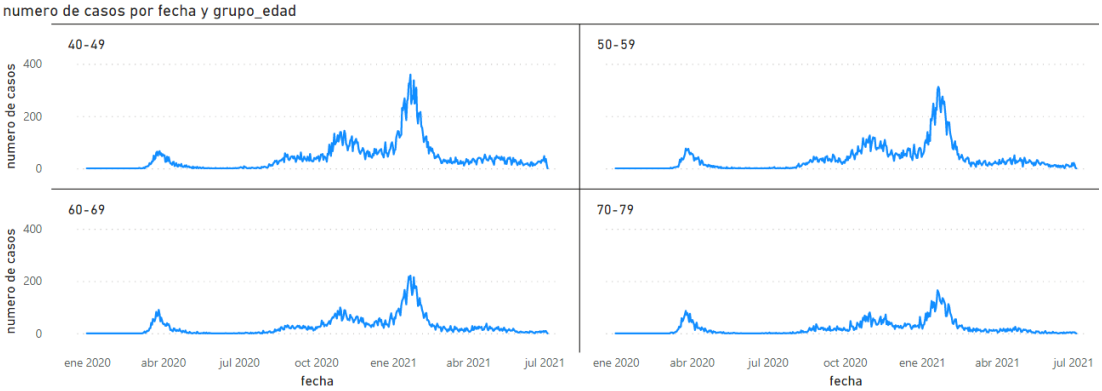

Figura 6.31: Informe que muestra, de forma diaria, el número de personas de entre 40 y 79 años que han dado positivo en COVID-19 a lo largo de la pandemia en Galicia

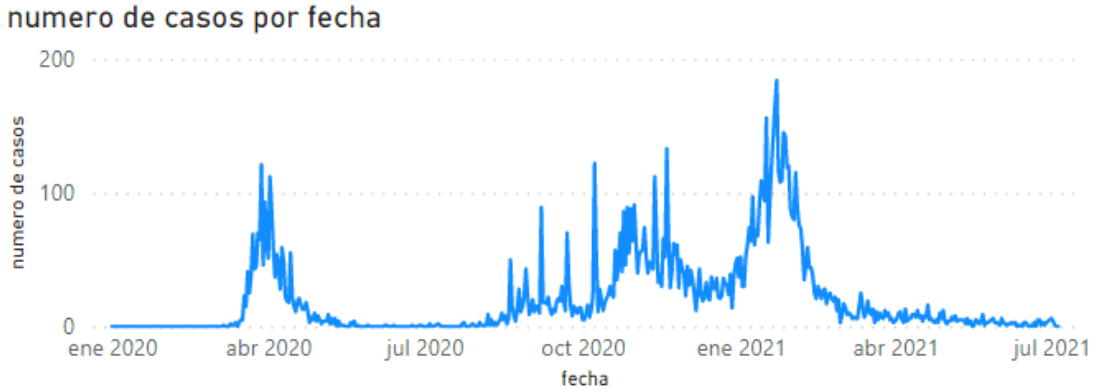

Figura 6.32: Informe que muestra, de forma diaria, el número de personas de 80 años, o más, que han dado positivo en COVID-19 a lo largo de la pandemia en Galicia

Los resultados de estos informes permiten apreciar que, las olas (picos) de marzo-abril

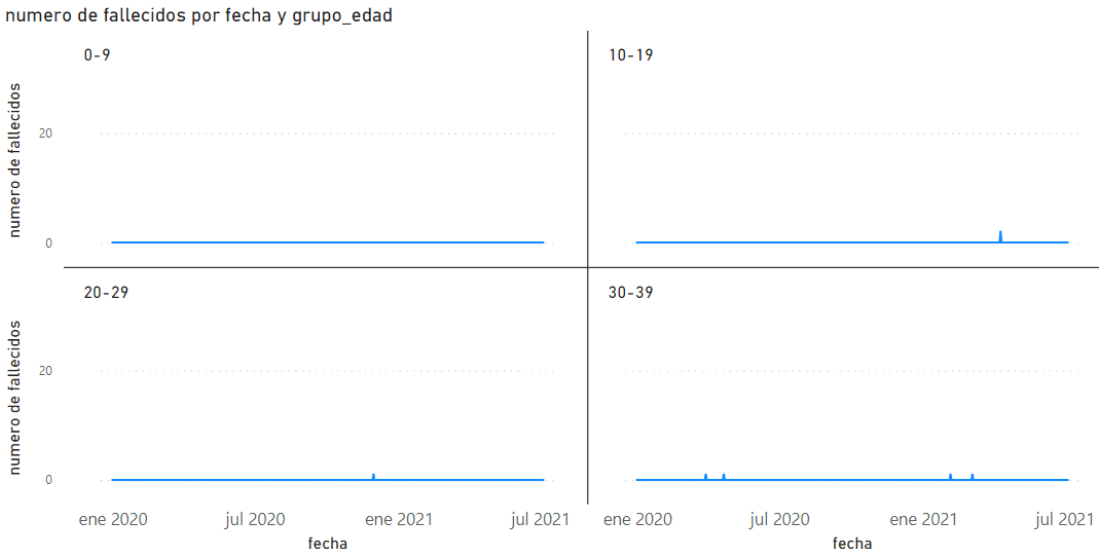

Figura 6.33: Informe que muestra, de forma diaria, el número de fallecidos menores de 40 años a lo largo de la pandemia en Galicia

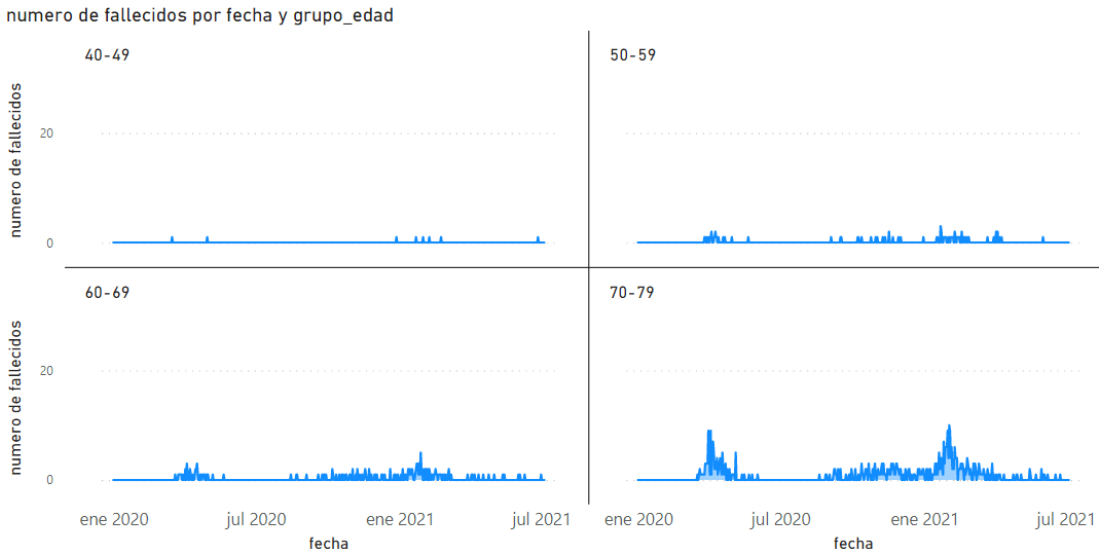

Figura 6.34: Informe que muestra, de forma diaria, el número de fallecidos de entre 40 y 79 años a lo largo de la pandemia en Galicia

(2020), octubre-noviemre (2021) y enero-febrero (2021) no han propiciado que haya un grupo de edad tenga muchos más casos que la media, aunque sí conviene destacar un par de anotaciones.

En la primera ola, la correspondiente a marzo-abril (2020), a pesar de que las personas de entre 40 y 59 años representaron un número de contagios a tener en cuenta, los grupos de edades que hacen referencia a los mayores, las personas con 60 años o más son las que re-

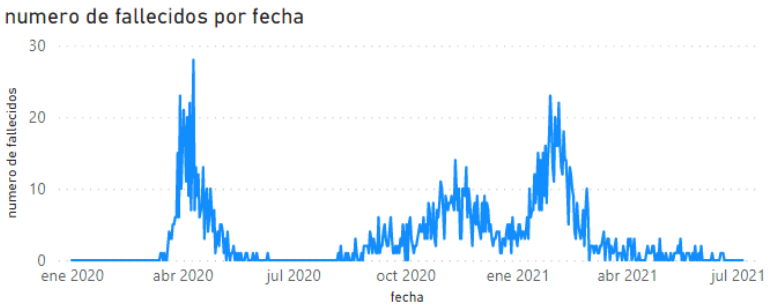

Figura 6.35: Informe que muestra, de forma diaria, el número de fallecidos de 80 años, o más, a lo largo de la pandemia en Galicia

presentan mayores números de contagios. De entre estos grupos de edades, los que acumulan ligeramente una mayor cantidad de contagios son los mayores de 80.

En la segunda ola, correspondiente a finales de octubre y principios de noviembre, se aprecia cómo comienza un proceso en el que los contagios pasan de ser más numerosos en las edades más avanzadas, a serlo en las edades medias. La diferencia no es mucha pero durante esta ola se aprecia más igualdad e incluso números superiores en los grupos de edad de entre 30 y 59 que en los de más de 60 años.

Es en la tercera ola (o pico), que comprende finales de enero y principios de febrero, donde los resultados son bastante diferentes en comparación con los de marzo-abril. Los grupos que representan las edades más avanzadas han pasado de ser los que acumulaban mayores números en cuanto a casos positivos, a ser un grupo que está en la media de contagios o un poco por debajo de ella.

En este tercer pico de la gráfica, los grupos de edades medios, desde 30 hasta 59 años, son los que más contagios han acumulado en estas fechas. Dentro de este conjunto de grupos de mediana edad, destaca un grupo en concreto, el de 40-49 años, el cual ha sido el protagonista de la tercera ola, llegando a alcanzar los 359 casos positivos nuevos el 22 de enero de 2021.

En cuanto a los más jóvenes, la primera ola prácticamente fue inexistente para ellos, no se ven casos hasta prácticamente finales de 2020 en los grupos de edades 0-9 y 10-19 años. Es en la tercera ola donde alcanzan los mayores números pero no son alarmantes en comparación con el resto.

Haciendo hincapié en los gráficos de número de fallecidos diarios, llama la atención que para los menores de 50 años, parece que el virus no ha tenido la letalidad que hemos visto. Son datos muy llamativos que permiten ver, con muchísima claridad, que el COVID-19 no es mortal para la gran mayoría de jóvenes y gente de edad media.

Es a partir de los 50 donde sí se comienza a atisbar un aumento de casos de fallecimiento pero también es bastante bajo. Es interesante ver cómo los grupos que durante la tercera ola han tenido una mayor cantidad de casos positivos, tengan una cantidad de muertes por día tan

baja. Esto reafirmar el hecho de que la gente que no pertenece a edades avanzadas, no corre un grave peligro al contraer el COVID-19. Es decir, el hecho de contagio para una persona menor de 50 significa que, con mucha seguridad, pasará la enfermedad y no fallecerá por la misma.

El claro despunte en número de muertes se encuentra en las personas mayores los 70 años. A pesar de que el número de muertes comienza a ser numeroso entre los 70 y 79 años, donde realmente se encuentran la mayor cantidad de casos es, con mucha diferencia, en las personas que han sobrepasado los 80 años. Es muy sencillo ver que la primera ola fue muy letal para esta gente, rebasando durante varios días la veintena de muertos diarios. Más adelante, en la tercera ola, a pesar de los casos positivos fueron más que en la primera, se aprecia una menor cantidad de fallecimientos diarios.

6.9 Restricciones

En este punto se trata de conocer cómo ha respondido la población gallega a las restricciones que se han ido imponiendo a lo largo de la pandemia. Se analizan las restricciones más interesantes y se explica qué es lo que han supuesto.

6.9.1 Uso de restricciones

En primer lugar se realizarán algunos informes que permitirán conocer qué restricciones fueron más y menos usadas, pero antes de comenzar con ello, conviene recordar lo mencionado en el Capítulo 4 (sección 4.2), donde se deja claro que la parte del estudio que trata de las restricciones, se basa en una estimación de cada provincia (o área sanitaria) gallega. Es decir, que se estima que en todos los lugares de la provincia se comparten las mismas restricciones en las mismas fechas.

Así pues, lo primero es conocer cuáles fueron las restricciones más utilizadas, ya sea porque son más útiles, porque son más fáciles de cumplir sin que suponga una alteración drástica en el día a día, porque son menos prohibitivas, etc. Y, de la misma forma, también se podrán distinguir aquellas que son menos usadas, o bien porque son demasiado prohibitivas y se utilizan únicamente en casos extremos, porque no son efectivas, porque se usan en unas condiciones muy específicas, etc.

Realmente a la hora de realizar este análisis hay que tener en cuenta que existen varios tipos de restricciones y, las referentes al tipo genérico, se deberían analizar por separado ya que suelen referirse a un estado o situación en el que se encuentra un lugar y suelen englobar otras restricciones. Por esta razón, se realizan dos informes por separado (Fig. 6.36), en uno se verán las restricciones más específicas y, en el otro, las restricciones de tipo genérico.

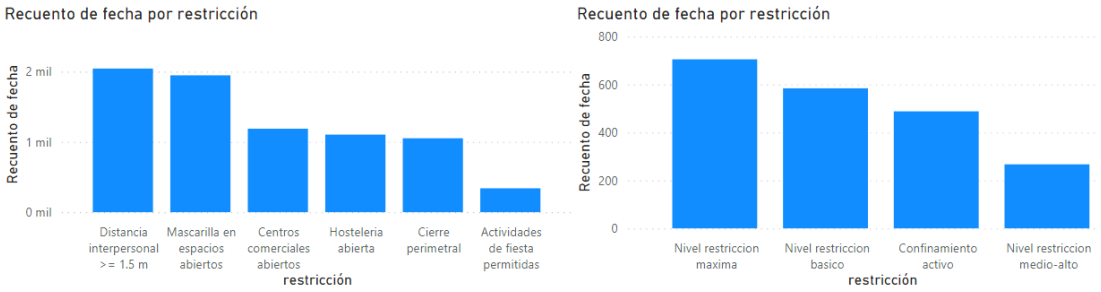

Figura 6.36: Informes sobre la cantidad de días que estuvieron activas las restricciones indicadas separando las de tipo genérico (derecha) de las demás (izquierda)

Tras realizar esta separación se aprecia con más claridad una diferenciación entre restricciones, donde se ve que la distancia interpersonal y el uso de mascarillas en espacios abiertos (y, evidentemente, cerrados) han estado activas durante todo el período que comprende el estudio (o prácticamente todo).

En cuanto a las restricciones de tipo genérico, resulta fácil distinguir qué niveles de restricción han sido más y menos usados. Por un lado están los niveles de restricción máximo y básico, que han sido los más utilizados y, por el otro, el nivel de restricción medio-alto, el cual no ha estado activo tanto tiempo como los anteriores.

6.9.2 Estado de alarma

A continuación gracias a la creación y visualización de informes, se busca conocer si el confinamiento o estado de alarma (Fig. 6.37) ha sido útil, ya que fueron meses muy duros y se intentará comprender si fue vital la implantación de esta restricción y si sus resultados han sido buenos. De esta forma se podrá comprobar cómo era la situación cuando se comenzó a implantar la restricción, cómo fue mejorando o empeorando dicha situación mientras estuvo activa y cómo ha afectado el levantamiento de la misma.

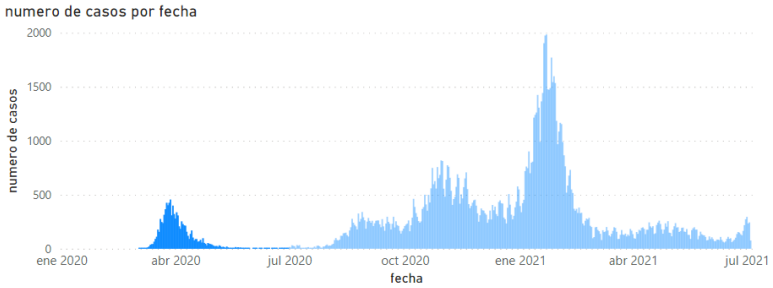

Figura 6.37: Informe que muestra los casos positivos diarios en Galicia durante el confinamiento/estado de alarma (parte sombreada del diagrama)

Gracias al diagrama, se puede apreciar cómo el confinamiento comenzó nada más co-

menzaron a aparecer los primeros casos, es decir, se actuó muy rápido. Además se ve cómo inicialmente hay un pequeño ascenso en el número de casos y se mantiene durante varias semanas, esto puede deberse a que, como sabemos, los síntomas del COVID-19 pueden llegar a aparecer a los 15 días de haber sido contagiado, por lo que es normal que a pesar de estar en estado de alarma, siguiesen apareciendo casos nuevos.

Tras ese período de tiempo en el que se mantuvieron constantes los positivos por día, comenzó una etapa en la que los contagios eran mínimos debido, por lo que parece, al estado de alarma.

Por último, no solo se debe mirar el período en el que el confinamiento estuvo activo, es interesante revisar cómo fue el comportamiento de la población los días o semanas posteriores al levantamiento de la restricción. Es evidente que el estado de alarma hizo que los contagios se frenaran en seco ya que durante meses éstos fueron mínimos y, una vez se levantó el estado de alarma, los casos positivos volvieron rápidamente.

En resumen, se podría decir que sí, definitivamente esta restricción cumplió su función de reducir al máximo los contagios y, muy probablemente, la situación actual podría haber sido muy diferente si no se hubiese decretado este estado en ese momento.

6.9.3 Cierre perimetral

Una de las restricciones más protestadas ha sido el cierre perimetral. Gracias a los datos de los que dispone el Data Warehouse, se pueden obtener los resultados en los que se muestra cómo ha afectado tanto la implantación de la restricción como su levantamiento (Fig. 6.38). Con ellos se intentará determinar si disminuyen los contagios, aumentan o si no varían.

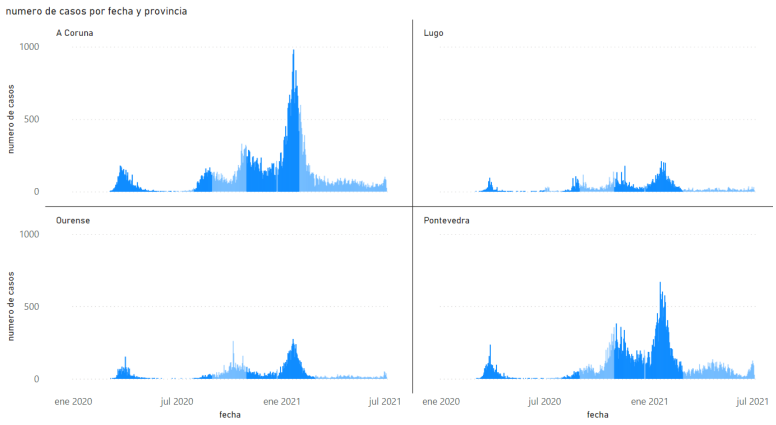

Figura 6.38: Informe que muestra el comportamiento de la población en base a la restricción de cierre perimetral)

Se sabe que durante algunos períodos de la pandemia, muchas de las zonas de Galicia se encontraban cerradas, es decir, no se podía ni entrar ni salir de ellas salvo excepciones. Esto

tenía como objetivo aislar los focos de contagios y que no se propagasen por zonas donde la situación era mejor.

En la gráfica se puede ver una primera etapa de cierre perimetral en la que no parece que sea una restricción de mucho peso y que tenga repercusión a la hora de detener los contagios pero, más adelante (sobre el mes de noviembre), se vuelve a implantar su uso tras unos meses de inactividad y parece que los casos comienzan a descender poco a poco. Este descenso puede deberse a que fue un momento en el que hubo bastantes contagios y el hecho de que se pusiese en cuarentena las zonas que eran focos de contagio, detuvo la propagación del virus a las zonas 'sanas'.

Como se ve, parece que la restricción estaba funcionando correctamente pero se aprecia un pequeño período a finales de diciembre en el que la restricción deja de estar activa. Efectivamente, a finales de diciembre, en fechas muy marcadas se levantó esta restricción y se permitió la movilidad de la población para pasar las fiestas navideñas con los familiares. Fueron muy pocos días, pero tras esta pequeña etapa en la que la gente podía moverse libremente por Galicia, llegó la ola más grande hasta el momento, cuadriplicando el número de contagios de marzo-abril (2020).

Cuando comienza la subida de esta tercera ola, la restricción está activa pero todo hace presagiar que los contagios fueron iniciados durante las fechas de levantamiento de restricción y fueron confirmándose en las siguientes semanas. A pesar de haber alcanzado el pico de contagios en enero-febrero, se ve una caída bastante brusca de casos positivos en COVID-19 que coincide con las fechas en las que el cierre perimetral estaba activo.

De esta forma, es bastante evidente que esta restricción ha sido muy útil y, bien utilizada, ha podido frenar los contagios de forma muy efectiva. A mayores se puede pensar que, si no se hubiese levantado el cierre perimetral en las fechas de final de año 2020, se podría haber impedido una subida tan grande de casos positivos al inicio del siguiente año que conllevaron a muchas muertes y hospitalizaciones. Por otro lado también hay que recalcar que se actuó bien quitando la restricción únicamente días señalados ya que si se hubiese esperado más tiempo a implantar de nuevo la restricción, se podría estar hablando de una ola mayor incluso.

6.9.4 Niveles de restricción

Para poder apreciar cómo de útiles han sido las restricciones de tipo genérico a lo largo de la pandemia, se decide obtener dos informes que representen sus números de contagios diarios desde inicios de 2020 hasta mediados de 2021 (Fig. 6.39 y Fig. 6.40). De esta forma se podrá ver cuándo se utilizan estas restricciones. si son efectivas, etc.

Estos dos informes, serán analizados de forma conjunta ya que ambas restricciones tienen mucho que ver entre ellas. Revisando en primer lugar los resultados que representan el gráfico referente a la restricción 'nivel de restricción máximo', se puede intuir que es una restricción

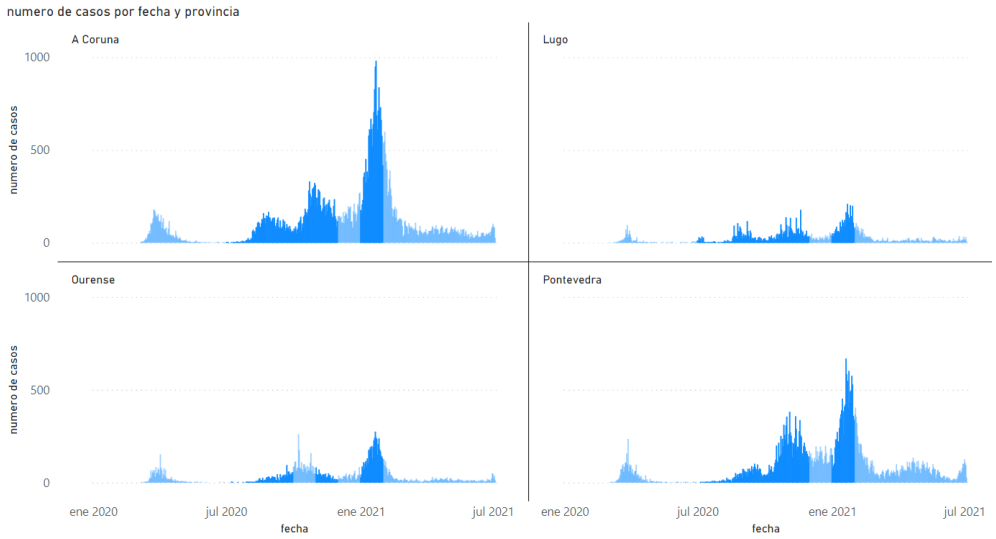

Figura 6.39: Informe que muestra el comportamiento de la población en base a la restricción de Nivel de Restricción Máximo)

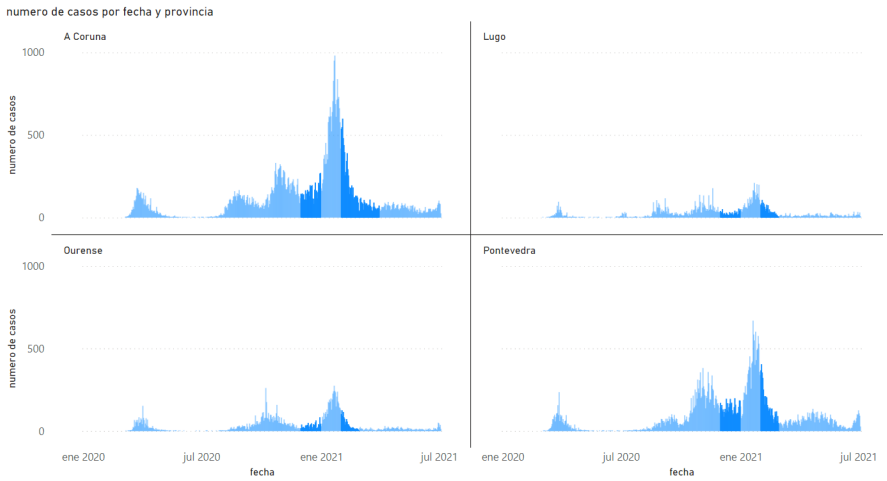

Figura 6.40: Informe que muestra el comportamiento de la población en base a la restricción de Nivel de Restricción Medio-Alto)

usada para realizar una desescalada. Es decir, fijándonos en su primera implantación, ésta es justo al terminar el estado de alarma se implanta esta restricción para que siga habiendo unas prohibiciones o pautas a seguir pero no tan restrictivas como el propio confinamiento. Con ello se puede ver cómo responde la población al levantamiento del estado de alarma. Los contagios comienzan a subir, llegando a alcanzar e incluso superar los números de la primera ola. En este período existen subidas y bajadas de contagios, por lo que parece que no es una restricción demasiado determinante, como puede ser el caso del confinamiento, en el que se aprecia claramente un descenso de casos de forma rápida y drástica.

Tras superar la segunda ola y comenzar a menguar los casos, esta restricción pasa a un estado de inactividad, dando paso a la aparición de 'nivel de restricción medio-alto'. Gracias al gráfico de la Fig. 6.40, se puede intuir que esta restricción es utilizada para continuar con ese proceso de desescalada que se comentaba, en el que, tras pasar por la restricción 'nivel de restricción máximo' y ver caer el número de contagios por día, se reducen todavía más las prohibiciones y pautas a seguir para intentar volver a un estado de normalidad.

Desafortunadamente los casos vuelven a aumentar en enero y es inviable continuar con las mismas restricciones, por lo que se vuelve al 'nivel de restricción máximo' para intentar frenar, junto con las demás restricciones, el número de contagios diarios. Parece, según se ve en el gráfico que la combinación de restricciones en ese período de tiempo están funcionando correctamente y comienzan a verse claros descensos en el número de casos positivos diarios.

Debido a estos descensos de contagios, se decide retomar el uso de la restricción 'nivel de restricción medio-alto' viendo que la situación comienza a mejorar.

Como se puede comprobar en estos informes complementarios, el 'nivel de restricción máximo' suele utilizarse como medida más drástica sin llegar al confinamiento o estado de alarma y su aparición es normalmente cuando se intuye que vendrá un aumento de casos positivos en COVID-19. Por otra parte, el 'nivel de restricción medio-alto' es utilizado como un intento de normalización más acentuado que el 'nivel de restricción máximo', y se utiliza cuando los contagios comienzan a alcanzar un nivel más o menos bajo. Además, por lo que se aprecia en el informe, parece que sí cumple su función y no da lugar a aumentos de contagios, is no que los contagios se mantienen o incluso bajan.

Para comentar todos los niveles de restricción, se puede comentar brevemente el uso del 'nivel de restricción básico' que, como es lógico, se ha utilizado en momentos en los que ha habido un número mínimo de contagios diarios en estos últimos meses, cuando comenzó el período de vacunación y la situación estaba mucho más controlada. El informe de la Fig. 6.41 respalda esta afirmación. Como anomalía, se puede apreciar que se intentó implantar de forma muy apresurada en Ourense pero, observando cómo crecía el número de contagios en esa zona, no duró demasiado tiempo.

6.9.5 Hostelería

La hostelería ha sido uno de los sectores más problemáticos de la pandemia y ha protagonizado muchas portadas a lo largo de esta, por ello se genera un informe (Fig. 6.42) para mostrar cómo ha afectado la prohibición o permisión de la apertura de bares, restaurantes, etc.

Este gráfico permite que nos demos cuenta de que es probable que la apertura de locales de hostelería haya sido el detonante de la segunda ola ya que a partir su apertura los casos subieron constantemente (a excepción de un par de días en septiembre de 2020) hasta propiciar

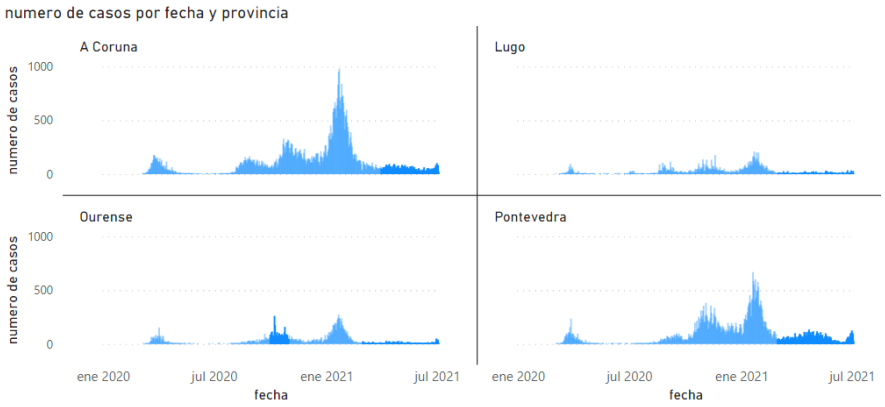

Figura 6.41: Informe que muestra el comportamiento de la población en base a la restricción de Nivel de Restricción Básico)

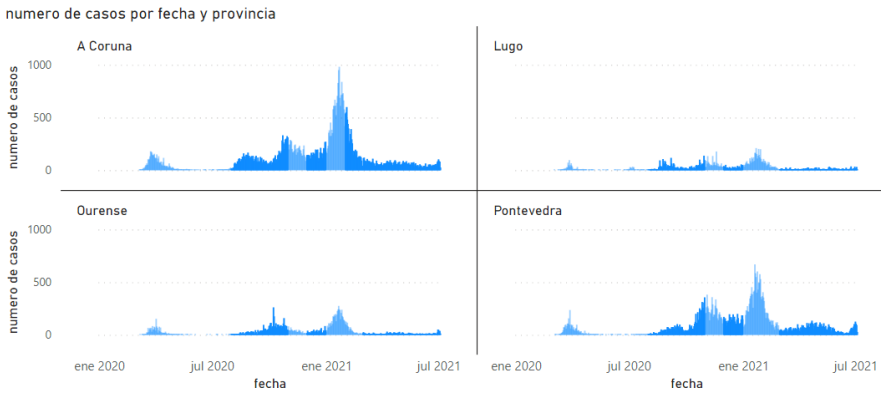

Figura 6.42: Informe que muestra el comportamiento de la población en base al cierre y apertura de la hostelería)

lo que se conoce como la segunda ola a finales de octubre y principios de noviembre (2020).

En las zonas más afectadas se prohibió su apertura y, una vez los casos volvieron a bajar, se volvió a intentar la apertura de la hostelería. En este segundo período se mantuvieron más los casos pero se siguen viendo pequeños aumentos de forma paulatina. Con la llegada del inicio de la tercera ola, viendo que los casos aumentaban drásticamente, se decide volver a prohibir la apertura de establecimientos de hostelería y unos meses después, cuando la situación está más controlada, sobre finales de marzo (2021), vuelve la hostelería a Galicia para quedarse hasta el día de hoy. En esta última etapa, los casos se mantienen al mínimo y esta es una de las razones por las que se ha mantenido abierta la hostelería.

6.9.6 Actividades de fiesta

Otra de las restricciones que sería interesante revisar sería la prohibición de actividades de fiesta. Sería interesante saber si, al permitirse estas actividades, el número de contagios

asciende debido a que la gente no toma las precauciones impuestas. De esta forma, se elabora un informe que permita ver los resultados de los contagios indicando cuándo estuvo activa esta restricción (Fig. 6.43).

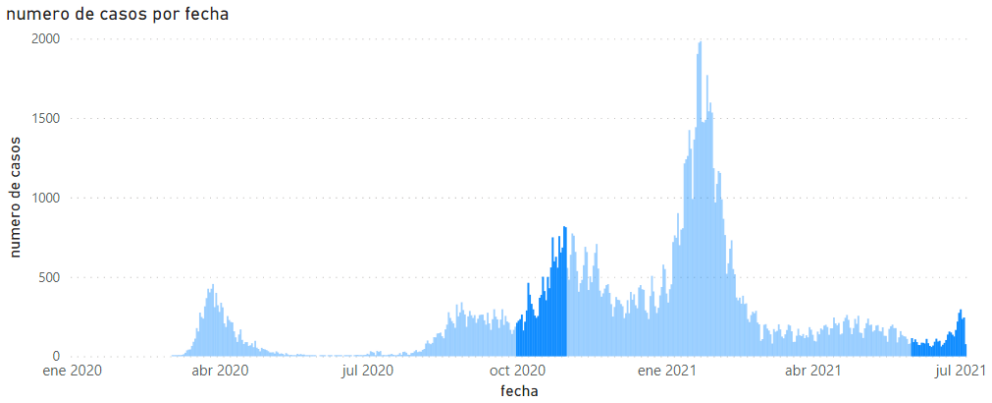

Figura 6.43: Informe que muestra el número de contagios en base a la permisión o prohibición de las actividades de fiesta)

Es muy sencillo ver que al permitirse las actividades de fiesta, los casos positivos en COVID-19 aumentan. Viendo estos resultados se puede decir que las actividades de fiesta son una de las razones por las que los contagios aumentan en Galicia. El claro ejemplo se encuentra en su primera permisión (octubre de 2020) en donde los casos comienzan a subir hasta alcanzar el pico de la segunda ola. Una vez se vuelven a prohibir estas actividades, los casos comienzan a descender, por lo que se puede intuir que uno de los culpables de la segunda ola ha sido la no prohibición de las actividades de fiesta.

6.9.7 Diferencia de impacto

A la hora de determinar qué grupos de individuos se vieron más afectados por alguna restricción, los estudios no dieron resultados útiles ya que, al vivir en una sociedad y ser el COVID-19 un virus con una alta facilidad de contagio, si un grupo determinado de personas aumenta los contagios, estos se repercuten casi al instante en el resto de grupos.

La mayor diferencia (sobre cómo ha afectado una restricción a cierto grupo de personas) se ha encontrado es con la restricción de la prohibición de las actividades de fiesta. Como se puede ver en la Fig. 6.44, el grupo con el aumento más notorio de casos ha sido el de las personas de entre 40 y 49 años. Si bien había otras restricciones activas en ese momento (como la hostelería abierta y el no cierre perimetral), el aumento no se produjo hasta pocos días después de la permisión de actividades de fiesta, por lo que todo parece indicar que esta fue la culpable del aumento (el detonante). Pero, como se acaba de comentar, puede que sea debido a otros motivos ya que el virus se propaga con mucha facilidad haciendo muy complicado este

estudio.

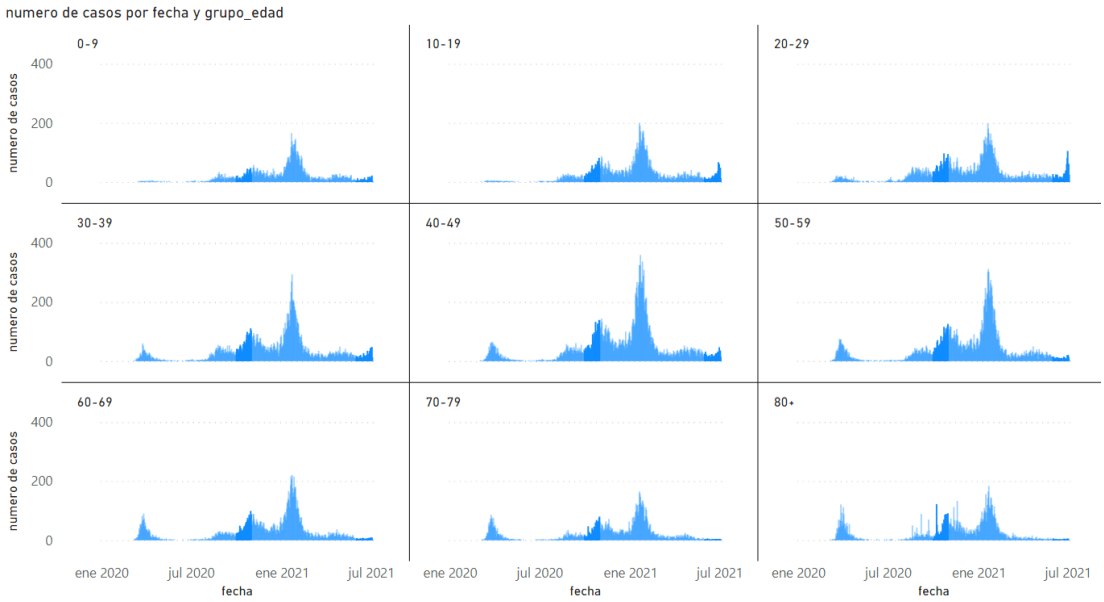

Figura 6.44: Informe que muestra el número de contagios por edades en base a la permisión o prohibición de las actividades de fiesta)

Capítulo 7

Conclusiones

Gracias a este trabajo, se han podido explicar algunos conceptos como lo que es un Data Warehouse, cómo es el proceso de construcción del mismo y cómo o para qué se suelen utilizar. Una de las conclusiones principales del proyecto ha sido comprender que los Data Warehouse son elementos muy importantes que se deben diseñar y mantener de forma correcta con el fin de que los datos estén listos para poder acceder a ellos y utilizarlos de la mejor manera posible para los fines que se deseen. Además de dejar clara su importancia, durante el proyecto se han conseguido ver y completar detalladamente los pasos que permiten el diseño, creación y mantenimiento del mismo.

También se ha visto de forma práctica cómo llevar a cabo análisis o informes utilizando el Data Warehouse para poder sacar conclusiones o entender mejor situaciones concretas. En este proyecto se realizaron numerosos informes que permitieron saber cómo era la situación en Galicia bajo la influencia del COVID-19. Se aprendió a importar los datos en la herramienta de Power BI, realizar ciertos modelados para poder realizar cruces entre tablas, generar informes de calidad, etc.

Y quizás, lo que más interés y en lo que se centra el proyecto es en saber cómo ha afectado el COVID-19 al territorio gallego y qué restricciones pudieron ser más o menos útiles, por ello ahora se realizará un resumen final con las conclusiones del estudio.

En primer lugar, hay que remarcar el importantísimo el papel de los datos e información a día de hoy ya que, en algo como la medicina que aparentemente no está muy relacionada, se nota cómo la recolección de datos durante la primera ola y la revisión de los mismos sirvió para saber cómo actuar frente a situaciones similares posteriores.

Sobre los informes vistos en el estudio, se pueden sacar varias conclusiones.

- Las provincias de Galicia han tenido diferentes resultados en los análisis que se han hecho y, en ellos, se aprecia claramente cómo A Coruña ha sido la que mayores números ha obtenido. No solo acumula el mayor número de contagios, también ha sido la que más muertes, hospitalizaciones e ingresos en UCI ha acumulado. En una segunda posición

estaría Pontevedra, seguida de Ourense y Lugo. De esta forma se concluye cuáles han sido las provincias en Galicia donde más personas padecieron el virus, fallecieron, se hospitalizaron e ingresaron en la Unidad de Cuidados Intensivos a causa del mismo.

- Otra situación muy distinta sería la de saber a qué provincia le ha afectado más el virus. Si se tiene en cuenta únicamente el recuento de contagios y demás parámetros, A Coruña y Pontevedra son las dos provincias a cuyos habitantes más afectó el virus (con diferencia), pero al tener en cuenta el factor población, el caso es diferente. Al realizar el mismo análisis utilizando el factor población, Ourense pasa a ser la provincia más afectada (junto con A Coruña), aunque los resultados de las cuatro provincias son muy similares. Esto es interesante puesto que permite ver con claridad la importancia de la población en este tipo de estudios.

- En cuanto a la cuestión de si el sexo es una característica importante a tener en cuenta para determinar si una persona puede morir tras contraer el COVID-19, gracias a los informes se ha podido determinar que sí, es un factor determinante.

  Esta afirmación no solo se ha deducido por el hecho de que los informes muestran que existe un mayor número fallecimientos en hombres, también es importante tener en cuenta el informe que representa la cantidad de hombres y mujeres que contrajeron el virus. Con ellos, se ve claramente que los hombres han acumulado menos contagios que las mujeres y, aún así, poseen un mayor número de fallecimientos, permitiendo afirmar que, con el conjunto de datos que se han hecho las pruebas, los hombres tienen más posibilidades de fallecer que las mujeres.

- La otra incógnita, con respecto a la mortalidad dependiendo del tipo de paciente, se encuentra en la edad. Se había comentado que sería interesante saber si la edad juega un papel decisivo a la hora de saber si la probabilidad de fallecimiento de una personas (por COVID-19) es mayor o menor. Con la ayuda de los informes vistos durante el proyecto, está claro que la edad juega un papel fundamental. Las personas mayores son mucho más propensas a fallecer por esta enfermedad que las personas de mediana edad y mucho más que los jóvenes, quienes prácticamente no han acumulado muertes en el período examinado.

- Otro de los puntos que se pretendía abordar era el conocer cómo afectaba el virus a los diferentes tipos de personas y, gracias a los informes generados, se ha sacado en claro que las personas mayores de 70 años son las que peor lo pasan con diferencia ya que no solo son las que más fallecimientos acumulan, también son los más hospitalizados.

  Por otro lado, el sexo masculino es el más presente en ingresos hospitalarios y, dentro de los propios hospitales, en la Unidad de Cuidados Intensivos. Esto es lógico sabiendo

que es el sexo con mayor mortalidad, o dicho de otro modo, es el sexo al que más le afecta (negativamente) el virus.

- Uno de los temas principales en este proyecto era el detectar cuales habían sido las restricciones más (y menos) usadas y las más (y menos) útiles. Como conclusión del estudio, se sacó en claro que las más utilizadas fueron aquellas prevenían de forma notoria el contagio del virus y que, además, no suponían una alteración excesiva en el estilo de vida. Es decir, guardar una distancia prudente con la gente y el uso de mascarilla.

  Por el contrario, las restricciones que menos se utilizaron fueron aquellas que supusieron un aumento de contagios cuando fueron implantadas. Por ejemplo, cuando se permitieron las actividades de fiesta, los contagios aumentaron drástica y rápidamente.

  Desde el punto de vista de la utilidad o efectividad de las restricciones, el estado de alarma (o confinamiento), que comenzó a inicios de 2020, fue una solución fantástica que hizo descender el número de contagios de forma estrepitosa hasta alcanzar unos mínimos sorprendentes. Frenó la primera ola de forma rápida y efectiva, sin duda, ha sido la mejor restricción en cuanto a detención de contagios y, en casos extremos, es la que se debería utilizar por sus resultados.

  Además del confinamiento, hay que destacar otras restricciones. La referente a actividades de fiesta ha sido sumamente importante ya que, a pesar de que los periodos de tiempo en los que esta restricción no estuvo activa son pequeños, se aprecia cómo los casos subieron de forma alarmante, haciendo que la prohibición de actividades de fiesta sea una restricción efectiva.

  Junto con la anterior, existe otra restricción cuya importancia se ve de forma especialmente clara en los gráficos de este proyecto, el cierre perimetral. Durante finales de 2020 el cierre perimetral estaba activo y los contagios estaban decayendo pero, en unas fechas señaladas de Diciembre de ese mismo año, se dejó de usar esta restricción, dando lugar de forma inmediata a la tercera ola, la que más contagios supuso con diferencia. Por ello es evidente que el cierre perimetral ha sido una restricción sumamente importante, útil y efectiva.

  En cuanto a los niveles de restricción, que indican el nivel en el que se encuentra una zona dependiendo de su situación (siendo el nivel máximo el menos permisivo y el básico el que más permisivo), por los gráficos parece que se han utilizado correctamente, usando éstos como pasos para alcanzar la normalidad de forma gradual.

  Se ve cómo el nivel máximo se utiliza después de la etapa de confinamiento para volver poco a poco a realizar las actividades pre-COVID y también se suele utilizar cuando los contagios son abundantes. El nivel medio-alto es utilizado cuando los casos comienzan

a alcanzar unos números menos alarmantes pero todavía a tener en cuenta. Por último, el nivel básico se suele usar una vez la situación está controlada hasta cierto punto, imponiendo algunas restricciones pero volviendo a una relativa normalidad.

En resumen, se puede concluir que las restricciones se han utilizado correctamente y de una forma más o menos eficaz exceptuando casos puntuales, dando como resultado la situación controlada en la que nos encontramos ahora.

- Otro de los puntos interesantes era saber qué restricciones habían afectado más a cada grupo de individuos con el fin de poder determinar cuáles usar para cada uno de ellos, pero los resultados no fueron demasiado satisfactorios ya que todas las restricciones han tenido prácticamente respuestas idénticas en los diferentes grupos de edades, sexo o localizaciones. Esto seguramente es debido a la facilidad de contagio del virus que hace que si un grupo de personas se contagia, propague el virus a los demás rápidamente haciendo que el impacto sea el mismo o muy similar en todos los grupos.

  Por ello, teniendo en cuenta los análisis sobre las diferentes restricciones vistos con anterioridad, independientemente de los tipos de individuos, se ha llegado a la conclusión que una buena combinación de restricciones, en el caso de que los contagios comenzasen a crecer de manera importante, sería: cierre perimetral, prohibición de actividades de fiesta y nivel de restricción máximo o medio-alto (restricciones que son muy efectivas pero no tan prohibitivas como el confinamiento). En caso de que los contagios estén más o menos controlados, se alternaría entre el nivel de restricción medio-alto y básico dependiendo de si están subiendo o bajando los contagios (poco prohibitivas y con buen funcionamiento con pocos contagios diarios), incluso se podría implantar el cierre perimetral en las zonas más afectadas para que no se propague y la prohibición de actividades de fiesta temporalmente.

  Si por el contrario hubiese situación extrema de contagios, lo mejor sería el estado de alarma para frenar en seco los contagios. En todos estos casos el uso de mascarillas y los 1,5 metros de distancia interpersonal sería obligatorio.

- Por último, a modo de predicción, tras analizar los datos históricos, nada hace prever que la situación vaya a empeorar a corto plazo y todo indica que se mantendrán los contagios de forma estable y baja, por ello, viendo cómo se ha procedido en casos similares durante la pandemia, todo hace presagiar que se podrá mantener el 'nivel de restricción básico'.

Líneas Futuras:

Por la naturaleza de este proyecto, se ha implantado un alcance que ha supuesto ciertas limitaciones. Si en un futuro se decidiese continuar con dicho proyecto y ampliarlo o mejorarlo. Algunas de estas ampliaciones o mejoras podrían ser las siguientes:

- Realizar el estudio a más bajo nivel y con una cantidad de datos mayor para poder determinar con más precisión el impacto del COVID-19 y el de cada restricción, además de utilizar otras medidas y/o agrupadores no usados en este proyecto.
- Se podrían realizar estudios de predicción aumentando también el alcance a España.
- Automatizar todavía más los procesos ETL.
- Se podría intentar profundizar más en un análisis de las restricciones de forma conjunta.

Glosario

aditiva Son métricas cuyos datos son no acumulados, es decir, datos únicamente del propio día. Por ejemplo el número de contagios que hubo por día. 28

COVID-19 El nombre hace referencia al virus (coronavirus) y el año en el que comenzó a expandirse por el mundo (2019). Es una enfermedad zoonótica (que puede transmitirse de los animales a los humanos) que produce síntomas respiratorios.. 1

CSV Los archivos CSV (Valores Separados por Comas) son archivos de texto con caracteres separados por comas permitiendo que los datos se guarden con una estructura de tabla.. 4

granularidad Indica el nivel de detalle con el que se guardará la información, en este caso, en el Data Warehouse. 14

proceso ETL Proceso que permite extraer datos de diferentes fuentes, transformarlos para que puedan cargarse en bases de datos de forma correcta. 4

pruebas PCR Las pruebas PCR (Reacción en Cadena de la Polimerasa) permiten detectar el ADN o ARN de un patógeno o de células anormales a las del individuo.. 23

Bibliografía

[1] "Almacén de datos," 2021, consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://www.wiley.com/en-us/Building+the+Data+Warehouse%2C+4th+Edition-p-9780764599446

[2] "Esquemas de estrella," consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-star

[3] "Esquema copo de nieve," consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-snowflake

[4] R. Kimball and J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Indianapolis, IN: Wiley, 2004. [En línea]. Disponible en: https://www.safaribooksonline.com/library/view/the-data-warehouse/9780764567575/

[5] F. Provost and T. Fawcett, *Data Science for Business.* Beijing: O'Reilly, 2013. [En línea]. Disponible en: https://www.safaribooksonline.com/library/view/data-science-for/9781449374273/

[6] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN: Wiley, 2013. [En línea]. Disponible en: https://www.safaribooksonline.com/library/view/the-data-warehouse/9781118530801/

[7] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, *The Data Warehouse Lifecycle Toolkit: Practical Techniques for Building Data Warehouse and Business Intelligence Systems*, 2nd ed., Tom, Ed. Wiley, 2008.

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," The CRISP-DM consortium, Tech. Rep., August 2000. [En línea]. Disponible en: https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf

[9] "Postgresql: la base de datos relacional de código abierto más avanzada del mundo," consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://www.postgresql.org/

[10] "Edición empresarial pentaho," consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho.html/

[11] "Obtenga claridad cuando más la necesita," consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://powerbi.microsoft.com/es-es/

[12] "Datos coronavirus," 2021, consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://coronavirus.sergas.gal/datos/#/gl-ES/galicia

[13] "Galicia-covid19," 2021, consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://github.com/lipido/galicia-covid19/blob/master/datos_nueva_web_sergas/CifrasTotais.csv

[14] "Covid-19," 2021, consultado el 15 de noviembre de 2021. [En línea]. Disponible en: https://rubenfcasal.github.io/COVID-19/