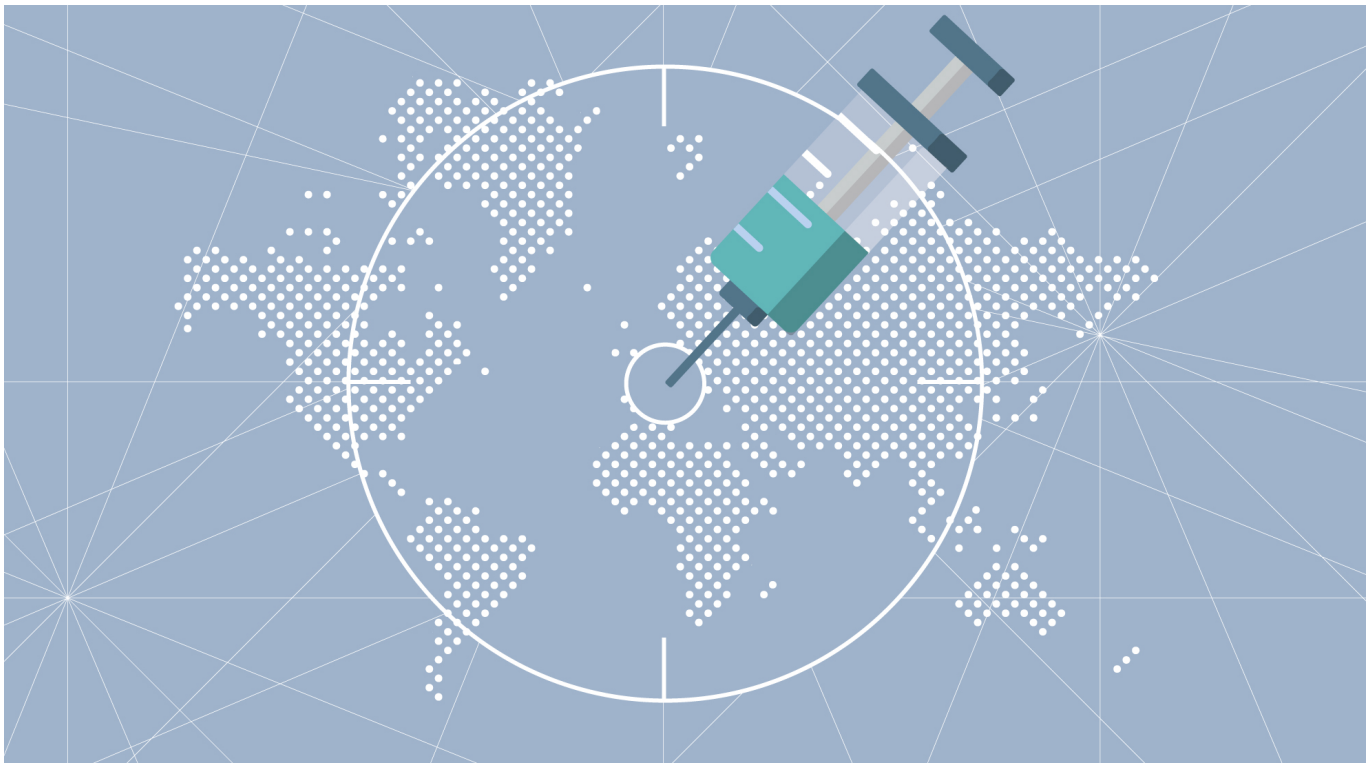
Universitat Oberta de Catalunya

M2.859 Visualización de datos

A8 - Creación de una visualización interactiva - PEC3

Marc Moreno Galimany
mmorenogali

Mayo, 2021



Índice

1 Dataset **1**

1.1 Descripción de las tablas 1

1.1.1 Dades diàries de COVID-19 per comarca 1

1.1.2 Registre de casos de COVID-19 a Catalunya per municipi i sexe . . 3

1.1.3 Vacunació per al COVID-19: dosis administrades per comarca . . . 4

1.2 Unión de las tablas 5

2 Mockup **6**

2.1 Realización del mockup 6

2.2 Elementos interactivos 9

3 Implementación **10**

3.1 Visualizaciones 10

3.1.1 Google Data Studio 10

3.1.2 Microsoft Excel 12

3.2 Comparativa 14

4 Funcionalidades de la visualización **15**

1. Dataset

Para la realización de esta práctica se han tomado tres conjuntos de datos de la página de datos abiertos de Cataluña `https://analisi.transparenciacatalunya.cat/browse?q=covid&sortBy=relevance`.

1.1 Descripción de las tablas

Los datasets son las tablas *Dades diàries de COVID-19 per comarca*, la tabla *Registre de casos de COVID-19 a Catalunya per municipi i sexe* y la tabla *Vacunació per al COVID-19: dosis administrades per comarca*. Los tres conjuntos son facilitados por el Departament de Salut, contienen datos de los distintos sistemas de información del Departament de Salut y del Servei Català de la Salut, y se han tomado en la actualización del 14 de mayo de 2021. `https://dadescovid.cat/documentacio`

En los respectivos enlaces se puede encontrar una descripción del dataset, que resumimos a continuación, junto con lo que se ha observado en la exploración:

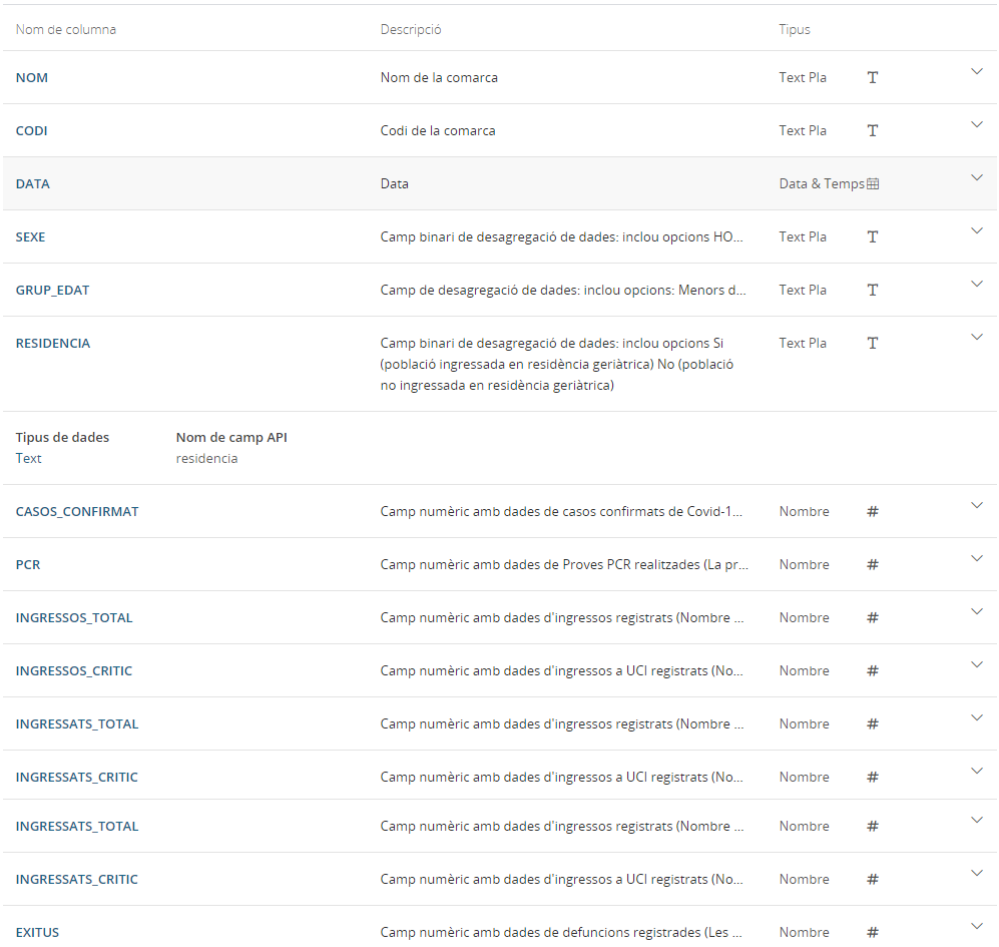
1.1.1 Dades diàries de COVID-19 per comarca

Esta tabla proporciona información para realizar un seguimiento de la evolución del COVID-19 en Cataluña.

Se incluyen los casos confirmados de Covid-19, pruebas PCR, ingresos, defunciones e institucionalización en residencias geriátricas por franjas de edad, sexo y comarca.

Para las personas en residencias no se muestran los valores para las variables sexo y grupo de edad. En ambos casos se muestra *Tots*.

En la misma página se muestran las características de los campos, así como el número de filas y columnas: 166000 filas y 13 columnas, que combinan datos numéricos, categóricos/textuales y campos de fecha, así como información geográfica. La descripción en detalle se puede encontrar en https://analisi.transparenciacatalunya.cat/Salut/Dades-di-ries-de-COVID-19-per-comarca/c7sd-zy9j



Nom de columna	Descripció	Tipus
NOM	Nom de la comarca	Text Pla
CODI	Codi de la comarca	Text Pla
DATA	Data	Data & Temps
SEXE	Camp binari de desagregació de dades: inclou opcions HO...	Text Pla
GRUP_EDAT	Camp de desagregació de dades: inclou opcions: Menors d...	Text Pla
RESIDENCIA	Camp binari de desagregació de dades: inclou opcions Si (població ingressada en residència geriàtrica) No (població no ingressada en residència geriàtrica)	Text Pla
Tipus de dades: Text	Nom de camp API: residencia	
CASOS_CONFIRMAT	Camp numèric amb dades de casos confirmats de Covid-1...	Nombre
PCR	Camp numèric amb dades de Proves PCR realitzades (La pr...	Nombre
INGRESSOS_TOTAL	Camp numèric amb dades d'ingressos registrats (Nombre ...	Nombre
INGRESSOS_CRITIC	Camp numèric amb dades d'ingressos a UCI registrats (No...	Nombre
INGRESSATS_TOTAL	Camp numèric amb dades d'ingressos registrats (Nombre ...	Nombre
INGRESSATS_CRITIC	Camp numèric amb dades d'ingressos a UCI registrats (No...	Nombre
INGRESSATS_TOTAL	Camp numèric amb dades d'ingressos registrats (Nombre ...	Nombre
INGRESSATS_CRITIC	Camp numèric amb dades d'ingressos a UCI registrats (No...	Nombre
EXITUS	Camp numèric amb dades de defuncions registrades (Les ...	Nombre

Figure 1.1: *Resumen de campos de la tabla. Fuente: Dades Obertes de Catalunya* `https://analisi.transparenciacatalunya.cat/`

1.1.2 Registre de casos de COVID-19 a Catalunya per municipi i sexe

En esta tabla encontramos, para cada día, municipio, sexo y procedimiento diagnóstico el número de casos identificados como positivos.

La fecha del caso es la fecha de inicio de síntomas, no la de la realización de la prueba diagnóstica.

En los casos en los que no se ha podido identificar el municipio de la persona, el valor de *MunicipiDescripció* es *No Classificat*, mientras que en municipios con una población inferior a 200 habitantes el valor es*Altres municipis*, con el fin de preservar la confidencialidad.

En este caso se indica que la tabla contiene 201000 filas y 11 columnas, que contienen, en general, variables binarias, identificadores, variables categóricas textuales, numéricas cuantitativas (número de casos) y un datetime. En la figura **??**

Nom de columna	Descripció	Tipus
TipusCasData	Data / Fecha / Date	Data & Temps
ComarcaCodi	Codi de Comarca / Código de comarca / Region code	Text Pla T
ComarcaDescripcio	Comarca / Region	Text Pla T
MunicipiCodi	Codi de municipi	Text Pla T
MunicipiDescripcio	Municipi / Municipio / Municipality	Text Pla T
DistricteCodi	Codi de districte	Text Pla T
DistricteDescripcio	Districte / Distrito / District	Text Pla T
SexeCodi	Codi de sexe (0-home, 1-dona) / Código de sexo (0-hombre...	Text Pla T
SexeDescripcio	Sexe / Sexo / Gender	Text Pla T
TipusCasDescripcio	Resultat	Text Pla T
NumCasos	Nombre de casos	Nombre #

Figure 1.2: *Resumen de campos de la tabla. Fuente: Dades Obertes de Catalunya* `https://analisi.transparenciacatalunya.cat/`

Para más información sobre los campos, se puede visitar la página https://analisi.transparenciacatalunya.ca de-casos-de-COVID-19-a-Catalunya-per-muni/jj6z-iyrp

1.1.3 Vacunació per al COVID-19: dosis administrades per comarca

En esta última tabla se detallan el producto administrado y el número de la dosis, el número de personas citadas en la fecha de referencia y el motivo de rechazo de la vacuna en los casos que corresponda por dia, comarca, sexo y grupo de edad.

Los datos hacen referencia a la población con derecho a recibir asistencia sanitaria de financiamiento público en Cataluña.

En los casos en los que no se ha podido identificar el municipio de residencia de la persona, el valor de los campos comarca y provincia es *No classificat*

En este conjunto de datos hay un total de 171000 registros y 12 columnas, que consta de campos textuales, numéricos y un campo de fecha y el campo de comarca que hace referencia a la posición geográfica. Entre ellos se encuentran variables binarias, categóricas, cuantitativas (como la fecha y el recuento) y cualitativas, como por ejemplo la dosis, o el fabricante de la vacuna.

Nom de columna	Descripció	Tipus
SEXE_CODI		Nombre #
SEXE		Text Pla T
PROVINCIA_CODI		Text Pla T
PROVINCIA		Text Pla T
COMARCA_CODI		Text Pla T
COMARCA		Text Pla T
EDAT		Text Pla T
DOSI		Nombre #
DATA		Data & Temps
FABRICANT		Text Pla T
NO_VACUNAT		Text Pla T
RECOMPTE		Nombre #

Figure 1.3: *Resumen de campos de la tabla. Fuente: Dades Obertes de Catalunya* `https://analisi.transparenciacatalunya.cat/`

Para más detalle sobre los campos de la figura **??** se puede visitar la página siguiente: https://analisi.transparenciacatalunya.cat/Salut/Vacunaci-per-al-COVID-19-dosis-administrades-per-c/cuwj-bh3b

1.2 Unión de las tablas

Para la unión de las tablas se han realizado los siguientes pasos:

- Agrupación de la tabla de municipios en comarcas.
- Normalización de los nombres de los campos.

El código correspondiente se puede encontrar en github `https://github.com/mmorenogali/pec3Visualizacion`.

2. Mockup

La creación del mockup funcional se ha realizado mediante la herramienta figma (`https://www.figma.com/` recomendada en la asignatura.

2.1 Realización del mockup

Antes de la realización del mockup se deben estudiar las preguntas que se quieren responder, analizar cuales de ellas se pueden responder realmente con los datos disponibles y a partir de estas, decidir el formato del dashboard.

Después de haber estudiado las distintas tablas y haber realizado el procesado descrito en el apartado 1, se conocen los datos disponibles en las distintas tablas y se hará una recopilación de la información que proporcionan los datos, a partir de la cual podremos desarrollar las preguntas y el esbozo. Esta información es, en registros diarios:

- El número de PCR realizadas y el número de casos confirmados.
- El número de ingresos hospitalarios.
- El número de ingresos críticos hospitalarios.
- El número de ingresados hospitalarios.
- El número de ingresados críticos hospitalarios.
- El número de defunciones críticos hospitalarios.
- La clasificación de los puntos anteriores según el sexo, la comarca, el grupo de edad, y la diferenciación de si pertenecen a residencia.
- El número de casos positivos según el tipo de prueba, sexo y comarca en Cataluña
- El número de dosis administradas por comarca, edad, sexo, dosis y fabricante en Cataluña.

Con estos datos disponibles se puede dar respuesta a algunas preguntas sobre el covid-19. En esta práctica se ha decidido por:

- ¿Cuál es el total de dosis administradas por comarca, provincia, sexo y edad en un intervalo de fecha determinado?
- ¿Cómo ha evolucionado la administración de las distintas dosis en la población catalana durante el intervalo determinado?
- ¿Hay correlación negativa entre el número de dosis administradas y el registro de ingresos, ingresos críticos y defunciones por Covid?
- ¿Cuál es la prueba que proporciona más casos positivos?
- ¿Cuál es la provincia que reporta un número de casos más elevado?

Una vez detalladas estas preguntas deberemos pensar en el diseño que proporcione esta información de la manera más clara posible. Por ejemplo, sería coherente visualizar el número de dosis y el registro de ingresos hospitalarios de forma alineada para poder comparar visualmente la evolución.

Se puede aprovechar la respuesta a la pregunta de *¿Cuál es la prueba que proporciona más casos positivos?* para mostrar también la evolución del número de casos positivos en el tiempo. Por ejemplo con un gráfico de barras apiladas que muestre la evolución en el tiempo.

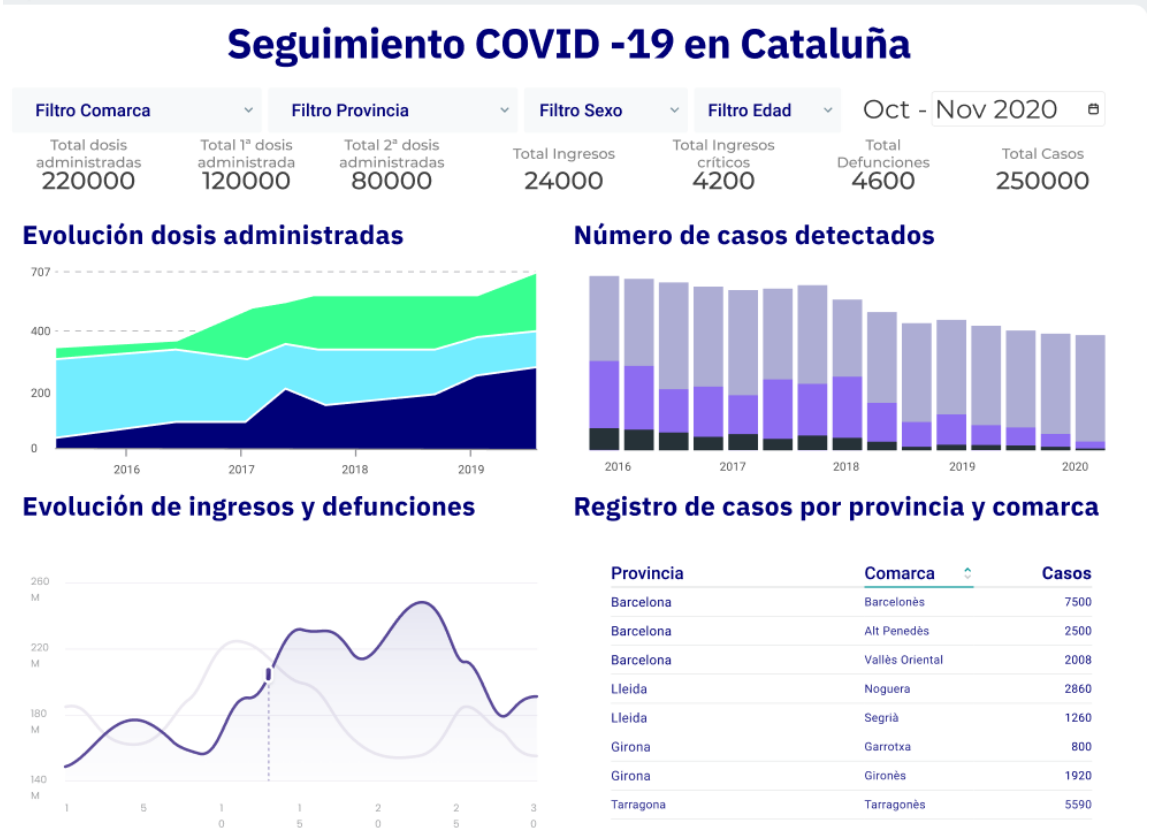


Figure 2.1: *Mockup en Figma. Fuente: Elaboración propia*

En la figura 2 se proporciona el diseño del mockup propuesto para la respuesta a las preguntas planteadas.

Es necesario mencionar que los valores no son reales, sino que son una muestra aleatoria generada para la realización del borrador.

El enlace a la visualización interactiva es el siguiente: `https://www.figma.com/file/KK2WgBN2yaETHauVg9COqC/PEC3-Mockup-mmorenogali?node-id=1253%3A275193`

2.2 Elementos interactivos

En la realización de la visualización se debe tener en cuenta el mantra de Schneiderman, y con este se creará el dashboard. De la visión general se debe permitir al usuario el filtrado y el zoom-in en los datos, y la muestra de datos bajo demanda.

Teniendo en cuenta el tipo de gráficos que se utilizarán los elementos más coherentes son:

- **Filtrado** Que permitirá al usuario la selección de los datos de interés.
- **Drill down** En las dimensiones temporales y geográficas. Esto le proporcionará el control sobre la granularidad de los datos, es decir, le permitirá realizar zoom-in en los datos de interés.
- **Hoovering**. Con esto se le proporcionarán al usuario los datos agregados en formato numérico, mostrando así los datos en el último nivel de demanda.

Veamos ahora las propuestas de visualización realizadas.

3. Implementación

3.1 Visualizaciones

3.1.1 Google Data Studio

En la figura 3.1 se puede encontrar una captura del dashboard realizado con Google Data Studio.

La intención de este dashboard es proporcionar todos los datos necesarios para responder a las preguntas iniciales planteadas además de proveer al usuario de datos bajo demanda.

Si realizamos una visión general del gráfico podemos observar, *a groso modo*, la **evolución de las dosis administradas mensualmente**, la **evolución de ingresos y defunciones mensuales**, el **número de casos detectados por tipo de prueba** y el **registro de casos por provincia y comarca**.

Sin embargo, rápidamente se observan datos más precisos. Es casi inmediato ver los valores numéricos que se proporcionan en la zona superior del tablero. Con estos datos se puede ver con más detalle el total de los distintos valores que se muestran numéricamente en el rango de fechas seleccionado en el filtro de la esquina superior derecha.

Estos filtros están destinados a proporcionar al usuario una visualización más personalizada, bajo demanda, que le permite seleccionar la(s) provincias, comarcas y sexo/género de la población que se quiera estudiar.

En el gráfico de la evolución de dosis administradas también se puede seleccionar la dosis que se quiere estudiar. Tanto si es la primera, la segunda como ambas. Por defecto se visualiza la segunda, ya que es la que, de alguna manera, asegura la inmunidad de la persona.

Siguiendo el mantra de Schneiderman, la visualización está pensada para que el usuario pueda seguir el orden de

- 1. Visión general
- 2. Zoom y filtrado
- 3. Detalles bajo demanda

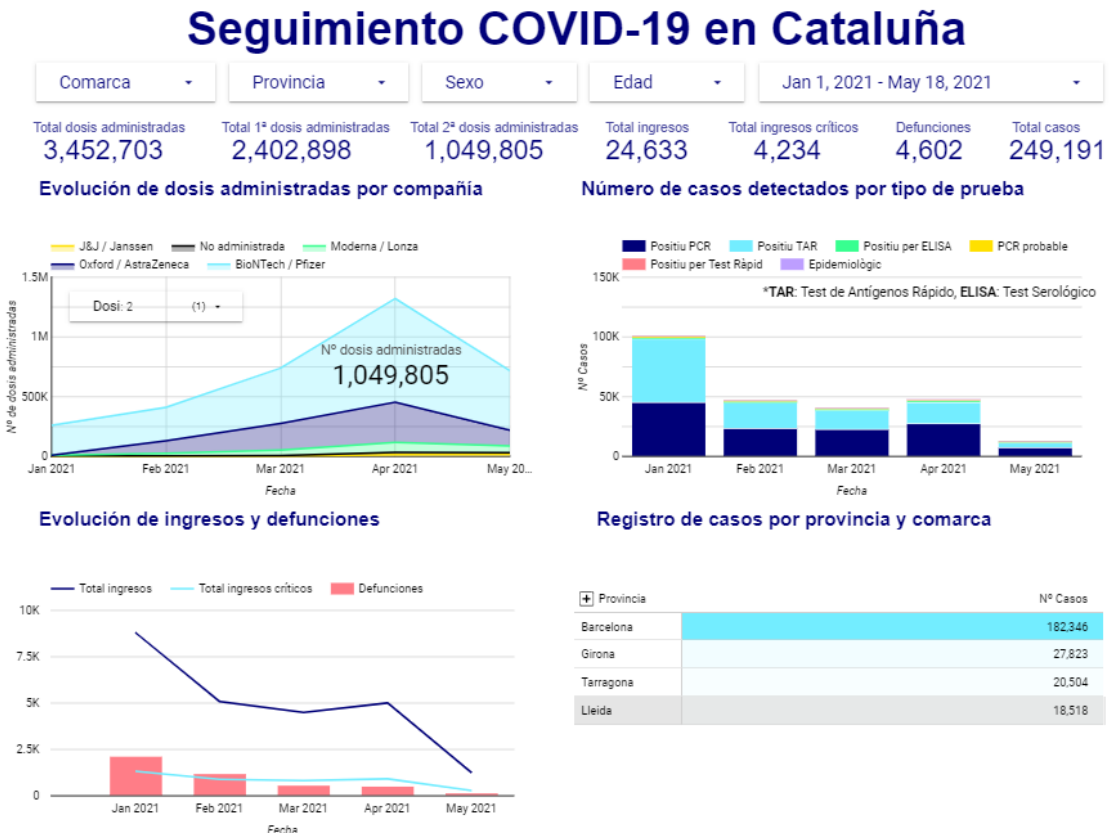


Figure 3.1: *Dashboard en Google Data Studio. Fuente: Elaboración propia*

La visión general y el filtrado ya se ha comentado. La parte del zoom es uno de los problemas encontrados con Google Data Studio, que no permite realizar un zoom con el ratón tal y como se realiza habitualmente. Sin embargo, todas las tablas permiten realizar un Drill Down en el campo de fecha para poder ver en detalle la evolución diaria de la lucha contra la pandemia en Cataluña.

Esto se consigue pasando el ratón por encima del gráfico. En este punto aparece una flecha hacia abajo que, en hacer clic permite realizar el drill down.

La tabla de **Registro de casos por provincia y comarca** permite realizar un drill down en la parte geográfica desglosando las provincias por comarcas. Esto es fácilmente realizable haciendo clic en el + que aparece al lado del título de Provincia. De este modo se despliega la tabla y se puede observar el número de casos por comarca.

El último punto del mantra de Schneiderman se consigue mediante el hoovering encima de los gráficos. Si se pasa el cursor del ratón por encima de cualquier gráfico, aparecen los datos detallados que permiten cerrar la visualización con el más alto detalle.

La versión interactiva de esta visualización se puede encontrar en el siguiente enlace:

`https://datastudio.google.com/reporting/bbfef14d-45ef-4e2f-b91d-dda4f4f31b96`

3.1.2 Microsoft Excel

Aunque no es específicamente un programa de visualización de datos, Microsoft Office Excel es una herramienta muy potente en todos los sentidos. Incorpora soluciones para la visualización de datos, que combinadas con las tablas dinámicas permiten realizar visualizaciones de gran calidad.

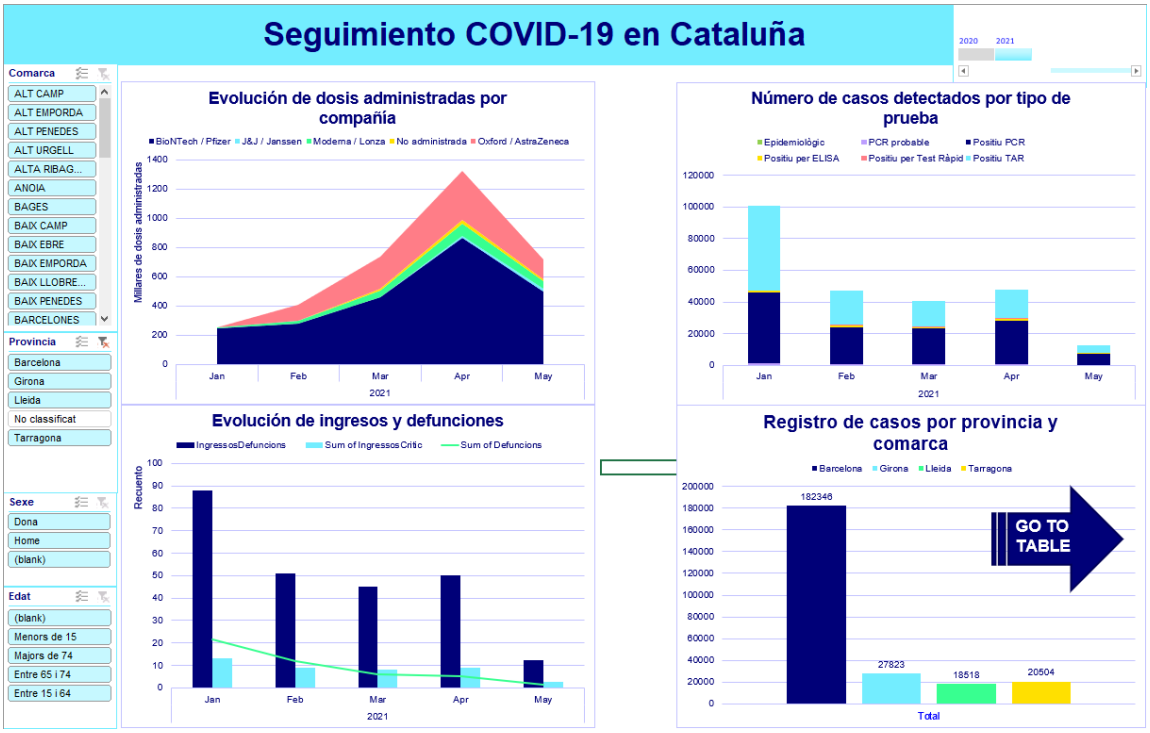


Figure 3.2: *Dashboard realizado en Excel. Elaboración propia.*

Para esta visualización se han tenido que combinar las tres tablas en una de sola. El código se puede encontrar también en el enlace de github mencionado anteriormente.

Se ha intentado seguir la misma estructura de datos que para el dashboard de Google Data Studio, pero en este caso los filtros se encuentran en la parte lateral del dashboard, ya que excel los proporciona con scroll vertical por defecto.

Estos filtros funcionan haciendo clic sobre los valores que se desean conservar o eliminar. Para realizar el reset se puede hacer clic en el símbolo de embudo de la parte superior derecha de cada grupo de filtro.

Por otro lado el rango de fechas se puede filtrar mediante un slide, una forma más visual pero menos precisa de definir el intervalo.

Por último comentar que una de las diferencias es que aplicar una tabla dinámica como la que se había generado en Datastudio no es posible, por lo que se ha utilizado un gráfico de barrras. Mediante la flecha que se puede observar en la zona superior derecha, se accede a la tabla dinámica donde se puede realizar el drill-down de cada provincia por comarca.

El dashboard se encuentra en el fichero "dashboardExcel.xlsx" que se adjunta en github.

3.2 Comparativa

Google Data Studio destaca la facilidad e interactividad con la que añadir los distintos gráficos en el dashboard, combinar distintas fuentes desde distintos orígenes, como por ejemplo subir un CSV, utilizar SQL, Google Analytics, Google Spreadsheets... Además no requiere de ninguna instalación, ni registro a parte del usuario y contraseña de la suit de Google.

Uno de los principales problemas encontrados con este software es que no permite una personalización en algunos aspectos. Por ejemplo, no se ha encontrado como rotar las etiquetas de los ejes horizontales para permitir que las fechas se muestren en vertical. Tampoco parece haber una manera que permita mostrar de una manera más clara o en otra posición el botón de drill down en los gráficos, lo que hace que los títulos se tengan que posicionar a una distancia considerable del gráfico, utilizando de forma poco eficiente el espacio.

También existe el problema de que DataStudio crea un archivo que vincula la tabla con el informe, lo que hace que para ver los posibles cambios se tenga que actualizar a mano este vínculo para refrescar la información.

Por otro lado Microsoft Office Excel permite la realización de un dashboard igualmente atractivo y customizable, aunque no es tan sencillo de utilizar. Requiere crear una tabla dinámica para cada gráfico, lo que puede resultar incómodo para dashboards con grandes cantidades de datos.

Sin embargo tiene la ventaja de ser un programa familiar, que la mayoría de usuarios utiliza y tiene instalado, y además funciona sin conexión a internet.

4. Funcionalidades de la visualización

Una vez realizados los dashboards se ve que gracias a las funcionalidades del software utilizado se puede ir más allá en la presentación de los datos. Por ejemplo, se ha podido aprovechar la pregunta de *¿Cuál es la provincia que reporta un número de casos más elevado?* para mostrar también en granularidad de comarca.

Se ha podido responder no solo a la evolución de los datos, sino a los datos totales en el período seleccionado. Por ejemplo, se puede estudiar la evolución de las dosis administradas en el tiempo, pero también se puede conocer el total de dosis administradas en el período, lo que enriquece la extracción de la información.

Del mismo modo el hecho de aplicar filtros permite al usuario responder a las preguntas no sólo de una forma general a nivel general, de provincia o de comarca, sino que le proporciona un nivel de personalización más elevado. De este modo puede responder a las mismas preguntas pero teniendo en cuenta una franja de edad, un género de la población, o una combinación de todas ellas.

También se puede extraer información más allá de las preguntas planteadas. Por ejemplo se observa que antes de finales de diciembre de 2020 no hay datos referentes a las dosis administradas, por lo que se podría deducir el momento de inicio de vacunación en Cataluña.