A8 (PEC3) Visualización de datos.

PEC realizada por: Manuel Ruiz Botella

Usuario UOC: manurubo

Máster universitario Ciencia de Datos.

1. Dataset

La A8 de la asignatura de Visualización de Datos se ha realizado con dos conjuntos de datos, aunque realmente uno de ellos es el principal, y el segundo conjunto de datos se ha utilizado simplemente para relacionar en una visualización de Datos el dataset 1 con los casos confirmados que aparecen en el conjunto 2, otorgando mayor profundidad a los conocimientos que se pueden obtener de visualizar el dataset 1.

El dataset 1 se titula "La societat catalana davant del Covid-19. Percepcions, estats d'ànim i preocupacions." y se puede acceder a él a través de: https://analisi.transparenciacatalunya.cat/Societat-benestar/La-societat-catalana-davant-del-Covid-19-Percepcio/qxjr-krv6. El autor del dataset es **Gabinet Ceres SL**, aunque su propietario es la **Generalitat de Catalunya** y fue actualizado el 10 de Mayo de 2021 aunque los datos se crearon el 19 de Mayo de 2020. El dataset se encuentra bajo la licencia "Llicència oberta d'us d'informació Catalunya".

Este dataset es un conjunto de datos que contiene los resultados de diferentes oleadas de encuestas que la empresa CERES realizó a la población catalán durante el confinamiento del COVID-19.

El dataset 2 se titula "Dades diàries de COVID-19 per comarca" y se puede encontrar en: https://analisi.transparenciacatalunya.cat/Salut/Dades-di-ries-de-COVID-19-per-comarca/c7sd-zy9j. El autor del dataset es **el Departament de Salut de Catalunya**, aunque su propietario es la **Generalitat de Catalunya** y fue actualizado el 21 de Mayo de 2021 (a diario) aunque los datos se crearon el 3 de Agosto de 2020. El dataset se encuentra bajo la licencia "Llicència oberta d'us d'informació Catalunya".

Este dataset presenta de manera transparente los casos, PCR, ingresos y defunciones de la población catalana en función de su edad, sexo, fecha y comarca.

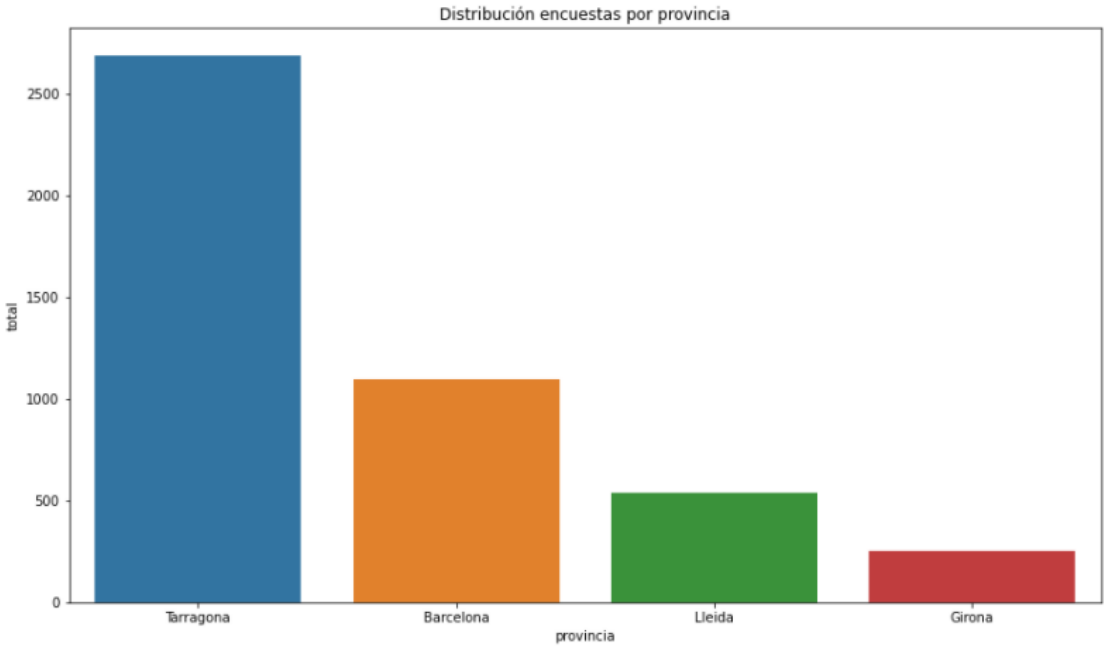
Todo el trabajo realizado para la PEC se encuentra en el github: https://github.com/manurubo/manurubo_a8_vd. Aquí se encuentra también un fichero de análisis de datos que se ha utilizado tanto para hacer un análisis exploratorio de los datos, como para generar los ficheros de datos que las visualizaciones realizadas utilizan.

Respecto al análisis exploratorio, nos vamos a centrar en el primer dataset. Este dataset contiene 4574 filas (encuestas) con 57 columnas (preguntas). **Como aspecto destacado, podemos ver estudiando los nulos del dataset que hay grupos de variables que tienen porcentajes de nulos similares**, por ejemplo, las variables que indican los días que ha salido una persona por tema (p_25_$razon) tienen un 15% de nulos, y las variables que estudian los sentimientos de las personas (p_23_$sentimiento) tienen todas entorno a un 42% de nulos.

	Nulos	PorcentajeNulos
p23_8_enutjat	1966	0.429821
p23_14_enfadat	1942	0.424574
p23_15_content	1941	0.424355
p23_11_molest	1940	0.424136
p23_13_intranquil	1938	0.423699
p23_12_animat	1938	0.423699
p23_4_melancolic	1938	0.423699
p23_10_apagat	1937	0.423481
p23_7_desanimat	1937	0.423481
p23_16_trist	1936	0.423262
p23_9_ansios	1934	0.422825
p23_6_optimista	1934	0.422825
p23_3_alegre	1930	0.421950
p23_5_tens	1926	0.421076
p23_2_irritat	1922	0.420201
p23_1_nervios	1919	0.419545
E1	1919	0.419545
E4	1918	0.419327
E3	1918	0.419327
E2	1917	0.419108
dies_sortida_m	713	0.155881
p25_99nc	702	0.153476
dies_sortida	702	0.153476
p25_4tabac	702	0.153476
p25_1Treballar	702	0.153476
p25_2comprar	702	0.153476
p25_3gos	702	0.153476
p25_97altres	702	0.153476
p25_6cura_persones	702	0.153476
p25_11passejarfills	702	0.153476
p25_5premsa	702	0.153476
p25_8compra_familiar	702	0.153476
p25_7metge	702	0.153476
p25_9alimentar	702	0.153476
dies_m	26	0.005684
p32_3	14	0.003061
p31_4	13	0.002842
p31_3	12	0.002624
p31_2	11	0.002405
edat_m	11	0.002405
p32_1	10	0.002186
p32_4	10	0.002186
p31_1	9	0.001968
p32_2	8	0.001749
id_onada	0	0.000000
onada	0	0.000000
sit_lab	0	0.000000
fills	0	0.000000
edat_rec	0	0.000000
sexe	0	0.000000
zones especifiques	0	0.000000
p46	0	0.000000
provincia	0	0.000000
mes_enquesta	0	0.000000
dia_Enquesta	0	0.000000
vpdef	0	0.000000
dies_rec	0	0.000000

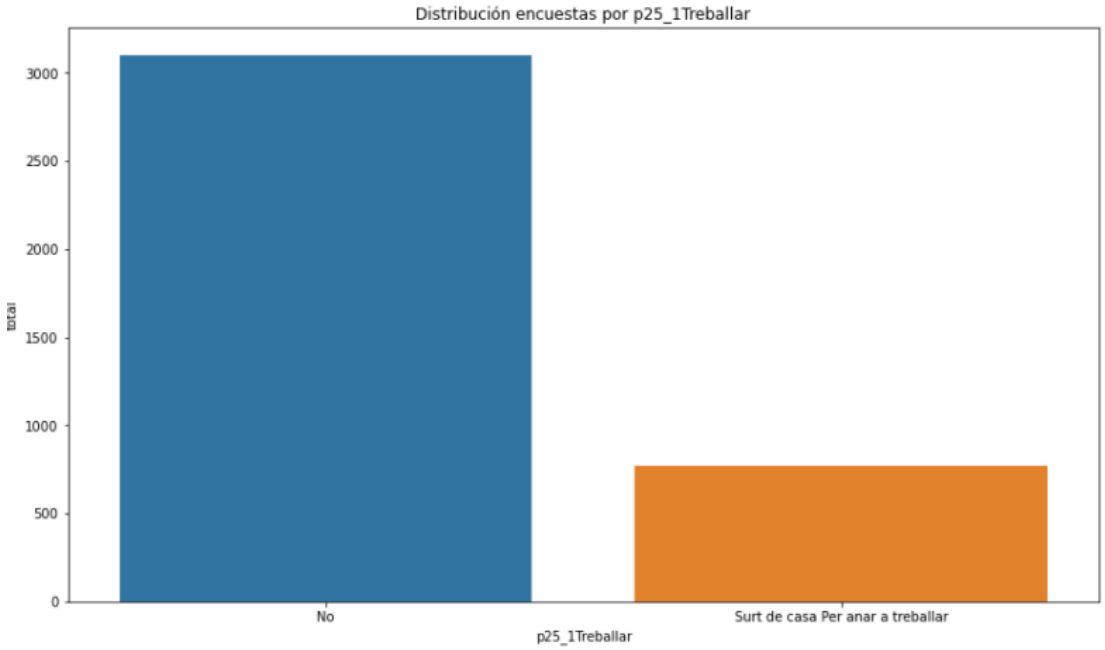
Es posible que esto sea porque hay oleadas donde no se realizaron estas, preguntas. Vemos que la variable dias_de_sortida tiene nulo en todas las instancias de la oleada 1, y las variables de sentimientos tienen nulos en todas las instancias de las oleadas 3 y 5. **Por tanto, concluimos que es cierto y que no en todas las oleadas se realizaron las mismas preguntas.**

Respecto a las variables que hay y sus tipos, encontramos que hay 29 variables categíoricas. Estas variables pueden ser de dos maneras, algunas son para conocer más el perfil del encuestado, como son aquellas que hacen referencia a su provincia o zona, su sexo, su edad agrupada, o si tiene hijos. Como ejemplo encontramos:



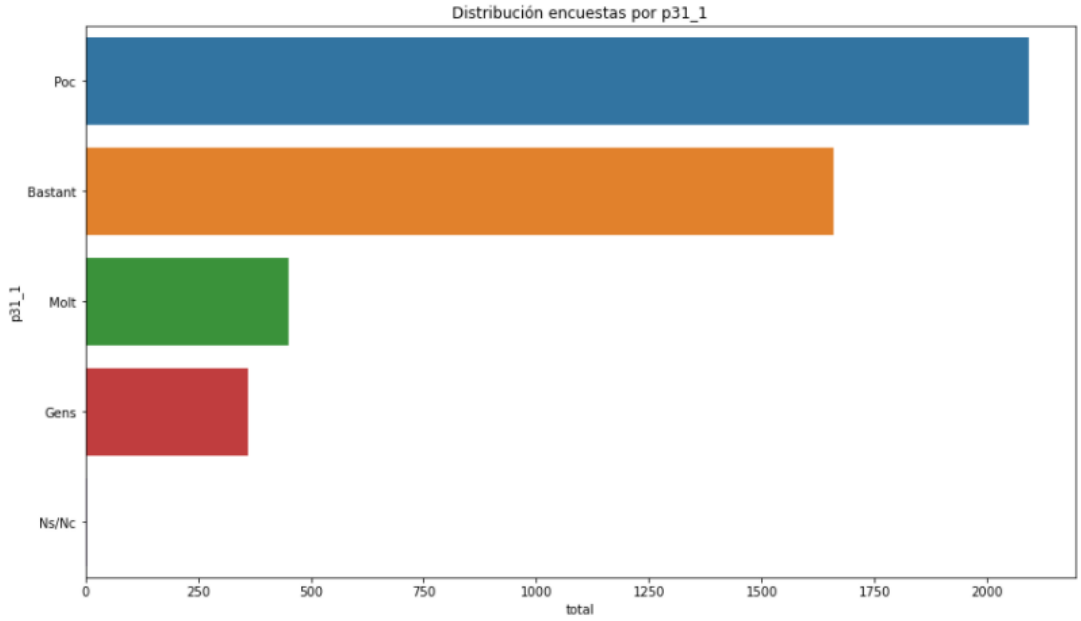
Donde vemos que la mayoría de encuestas fueron en Tarragona.

Mientras que hay otras variables categóricas que son preguntas de como está pasando la pandemia el encuestado. Hay dos grandes grupos de estas variables categóricas. Por una parte, aquellas de p_25_$motivo, que son categóricas respecto a los motivos por los que ha salido de casa el encuestado, como ejemplo tenemos:



Donde vemos que hay entre 500 y 1000 personas que salieron de casa para trabajar.

Y otras aquellas de p_31 y p_32, que son respectivamente sobre si el encuestado está preocupado por economía o por salud en diferentes ámbitos como son el familiar, el personal, el de amistades o el de la sociedad. Un ejemplo es:



Donde vemos que para la pregunta p_31_1 que hace referencia a la preocupación por la salud propia, la gente estaba mayormente poco preocupada, seguido de bastante, pero no estaban en los extremos normalmente.

Respecto a las variables numéricas hay 24 floats y 4 enteros, las 24 float hacen referencia unas pocas a la edad en número, los días que estima la persona que va a durar el confinamiento en número o los días de salida en número, y la mayoría son p_23_$sentimiento, que hacen referencia a la puntuación para el sentimiento en concreto o E(1-4) que es un resumen de varios sentimientos juntos. Cada una de estas variables de sentimiento puede tener un valor entre 1 y 10, siendo 10 que el encuestado siente mucho este sentimiento.

Las 4 enteras son el identificador de la persona, de la oleada de encuestas (se ha visto que hay 5 oleadas en el análisis exploratorio), el día y el mes de la encuesta, que vienen separados. **Como aspecto interesante, y con el objetivo de unir los datos a posteriori con los casos diarios del dataset 2, se ha generado una variable Fecha, con tipo fecha, a partir de este día y mes.**

Respecto al análisis exploratorio del dataset 2, apenas se ha realizado porque nos interesan solo los datos de casos por día, es decir, la variable casos confirmados y la data. No obstante, el dataset consta de más de 165 mil filas (cada día más) y 11 columnas, que son el nombre de la comarca, el código, la fecha que posteriormente se transforma a formato fecha, el sexo, el grupo de edad, la residencia, los casos confirmados, las pcr, y los ingresos(4 variables) y muertos. **En definitiva, lo interesante para realizar la última visualización de la actividad, ha sido ver que se puede agrupar los datos de cada comarca para tener los casos totales en Cataluña por día y seleccionar solo aquellas fechas que nos interesan por el dataset 1.**

	CASOS_CONFIRMAT	Fecha
0	2311	2020-03-23
1	1631	2020-03-20
2	957	2020-03-22
3	1059	2020-03-21
4	1921	2020-03-24
5	1137	2020-03-28
6	1947	2020-03-30
7	1624	2020-03-27
8	1463	2020-03-31
9	1001	2020-03-29
10	746	2020-04-04
11	1384	2020-04-03
12	562	2020-04-05
13	1173	2020-04-07
14	1417	2020-04-06
15	525	2020-04-12
16	1319	2020-04-09
17	1092	2020-04-10
18	1171	2020-04-14
19	798	2020-04-11
20	758	2020-04-13
21	1248	2020-04-15
22	456	2020-04-29
23	334	2020-04-27
24	436	2020-04-28
25	314	2020-04-30
26	160	2020-05-02
27	214	2020-05-01
28	75	2020-05-03

2. Mockup

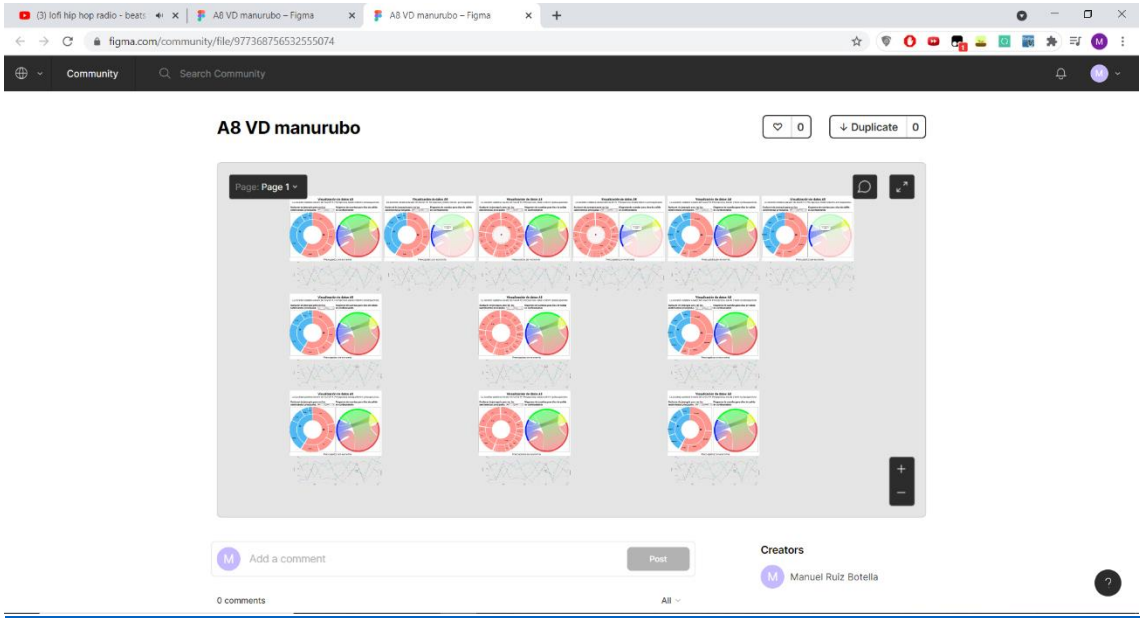
El Mock up se ha realizado en la herramienta Figma y se encuentra:

- Para editar en: https://www.figma.com/file/IoUGPQIeOWtLcq93oHUueC/A8-VD-manurubo?node-id=0%3A1
- Para editar publicado como parte de la comunidad en: https://www.figma.com/community/file/977368756532555074
- Para visualizar el prototipo en la web de figma: https://www.figma.com/proto/IoUGPQIeOWtLcq93oHUueC/A8-VD-manurubo?node-id=3%3A2&scaling=min-zoom&page-id=0%3A1
- Y para visualizar el prototipo embebido en un html en el github pages: https://manurubo.github.io/manurubo_a8_vd/mockup.html

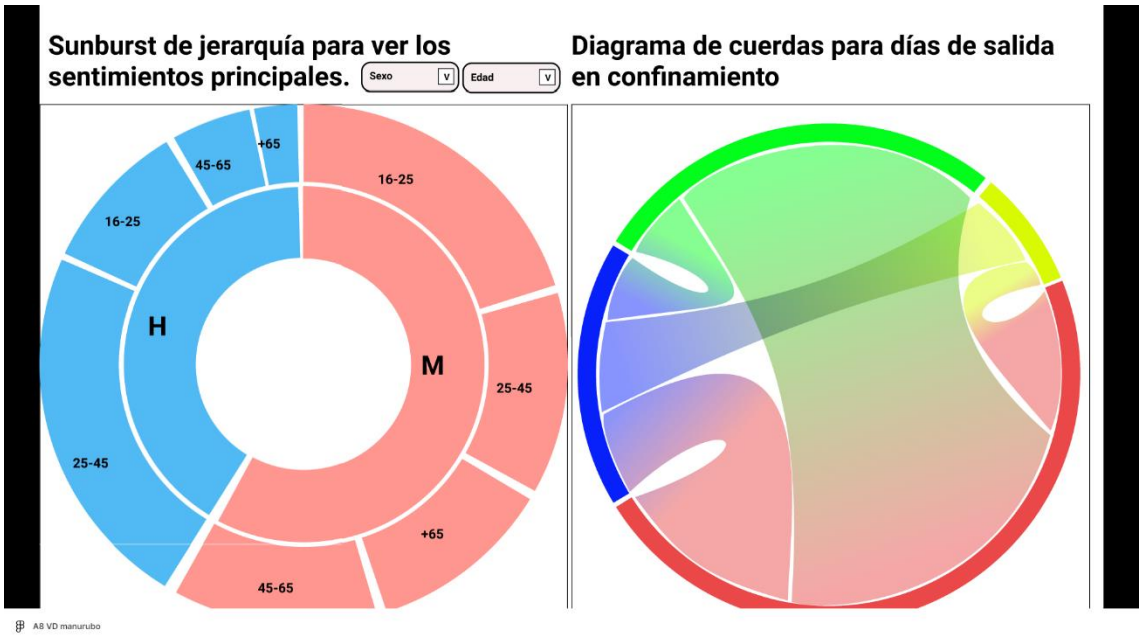
El prototipo realizado es una simple prueba de concepto para plantear las visualizaciones que se realizan más adelante. Se ha usado para diseñar la disposición de los gráficos en la visualización, para identificar que visualizaciones se podían realizar con los datos que tenemos y para idear los elementos interactivos de las visualizaciones y como funcionarán estos en la visualización.

Eso sí, la interactividad de estos elementos se ha realizado en parte, como ejemplo, no se ha hecho para que cada parte sobre la que se deberá clickar en la visualización real se pueda clickar, sino solo en una parte como ejemplo.

Si vemos el segundo link, podemos ver como se ve el diseño del mockup en todas sus etapas. Vemos que tiene 12 pantallas diferentes, partiendo de una inicial, debido a los elementos interactivos.

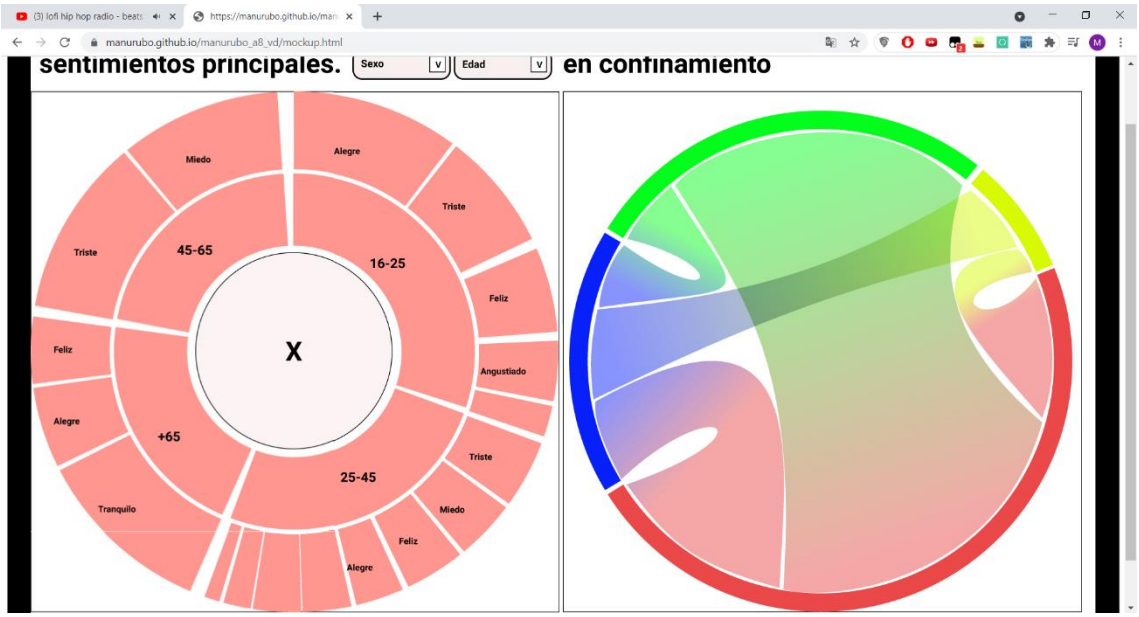


Si abrimos el mockup en https://manurubo.github.io/manurubo_a8_vd/mockup.html nos encontramos con 2 gráficas iniciales (y una abajo con scroll), **que permiten un overview de la visualización**:

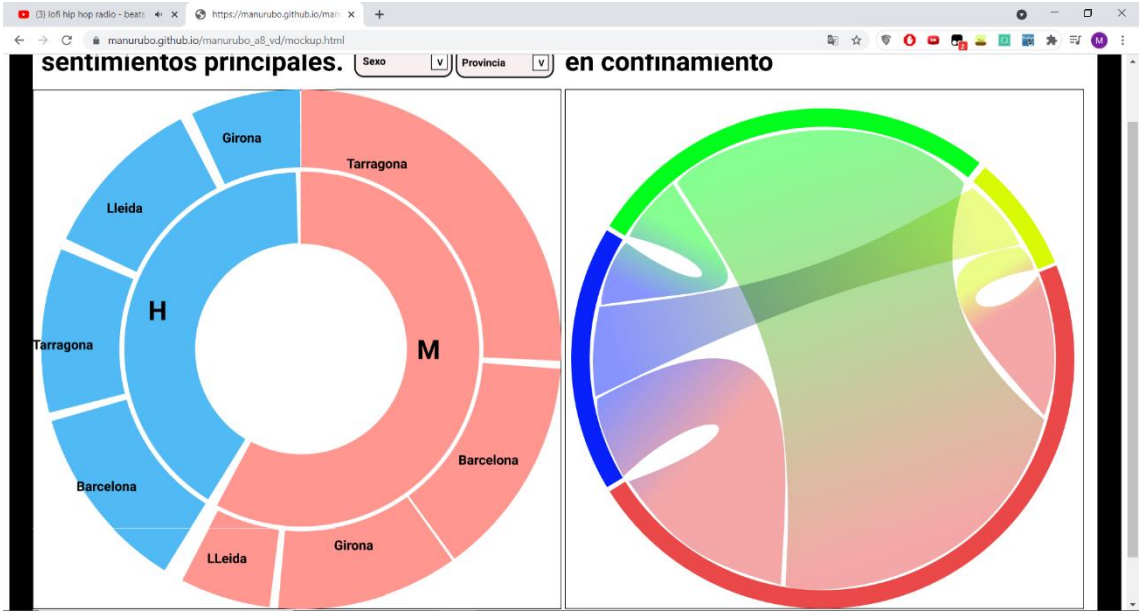


Que son un sunburst y un diagrama de cuerdas. El sunburst se plantea que muestra solo dos jerarquías en pantalla, en este caso se puede ver que se han seleccionado en los botones las jerarquías Sexo y Edad, por lo que se muestran estas en el Sunburst**. El objetivo será poder ver**

los sentimientos principales de cada encuestado (p_23_$sentimiento) como tercer nivel del sunburst agrupados por la selección jerarquica realizada, por lo que para visualizar los sentimientos principales será necesario interactuar con el gráfico. Por tanto, el sunburst incorpora dos elementos interactivos, por una parte, **clickando cerca de la sección de M (sexo = M), se puede filtrar los datos y acceder al siguiente nivel:**

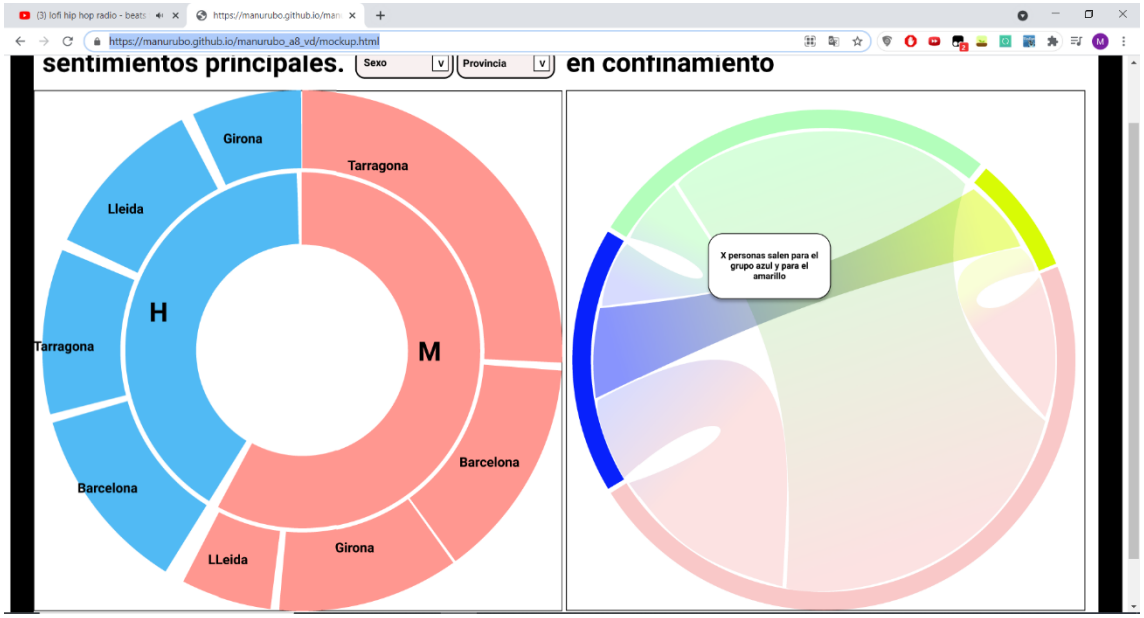


En este siguiente nivel vemos ya que se aprecia en el circulo interno las edades que se veían anteriormente en el círculo externo, y en el externo los sentimientos principales. Lo que cumple el objetivo de la visualización que queremos hacer, que es ver los sentimientos principales de los encuestados en una jerarquía. **Si ahora, clickamos en la X central, se vuelve al inicio del mockup. Una vez ahí, clickando en el segundo botón, se deberá de poder cambiar la jerarquía, por lo que se ha implementado en el mockup que se cambia el valor del botón a "Provincia", y se muestra el sunburst con "Sexo" y "Provincia" como jerarquía, es decir, se selecciona los datos a visualizar:**



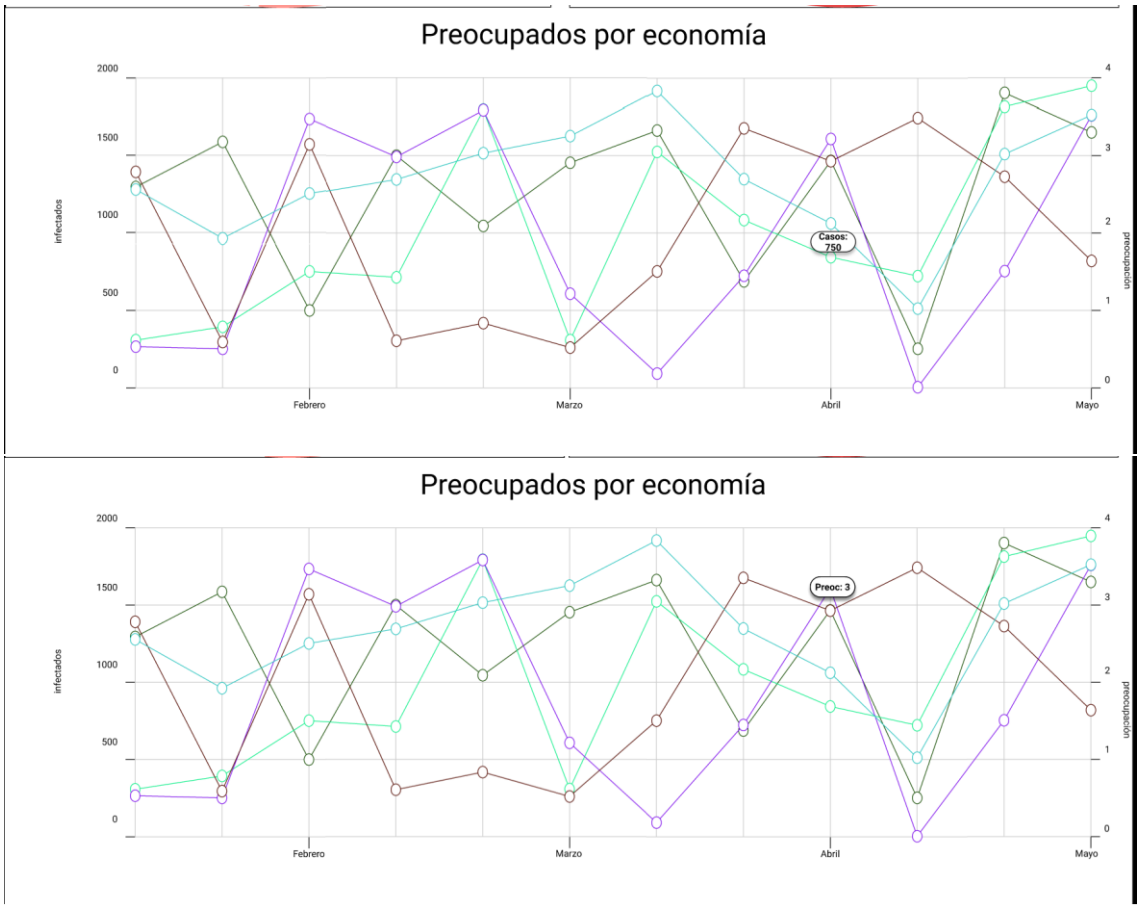
Una vez aquí, podemos estudiar ya las siguientes visualizaciones de datos**, la de la derecha es un diagrama de cuerdas, que se plantea que sirva para visualizar las relaciones entre los**

motivos por los que salen los encuestados del confinamiento. Como aspecto interactivo, permite ponerse sobre o cerca del link de azul a amarillo, de esta manera se destacará este link respecto al resto **y se proporcionan detalles bajo demanda de la relación**:



Por último, **como tercera visualización de datos, se va a buscar relacionar, la preocupación de las personas encuestadas por la economía (p_32_(1-4)) con los casos confirmados de covid-19 a lo largo del confinamiento**, que se encuentran en el dataset 2. Esta visualización se puede ver en el mockup si se hace scroll para abajo.

Esta gráfica del mockup tiene dos ejes, para poder comparar la escala de la preocupación con la escala de infectados a lo largo de los meses. **Como interacción, de nuevo destaca que se proporcionan detalles bajo demanda**, en esta ocasión hay dos puntos que permiten mostrar su valor, aunque en la realidad todos mostrarán su valor si pasamos sobre ellos.



En definitiva, vemos como el mockup desarrollado íntegramente en Figma con tutoriales que permite la herramienta y subido a github pages es un concepto que no solo idea las visualizaciones a plantear, sino que permite que quién tiene que desarrollar la visualización final comprenda que interactividad tiene que tener la herramienta.

3. Implementación de la solución interactiva

La visualización de datos que incorpora las tres visualizaciones presentadas en el mockup se encuentra en: https://manurubo.github.io/manurubo_a8_vd/manurubo_a8_vd.html

La visualización pretende ser un dashboard de las 3 visualizaciones donde se presenta el dataset y para cada visualización se explica un poco el motivo por el que se realiza la visualización y como se obtienen los datos en las encuestas realizadas**. Este dashboard se encuentra en un fichero html, que tiene las gráficas insertadas realizadas por otras herramientas, código html y Bootstrap para dar estructura a la web, un poco de javascript y jquery para interactuar con la tercera visualización y un fichero de css para dar estilo extra al html.**

Es importante mencionar, que en el mismo archivo de exploración que se encuentra en el repositorio github se generan los datasets derivados para realizar las visualizaciones de datos.

Tanto la visualización del sunburst como el diagrama de cuerdas se han realizado en Observable, a partir de ejemplos ya realizados, partes de código de otras visualizaciones, y partes de código desarrollado por el autor (yo). El sunburst se puede encontrar en solitario con el código en: https://observablehq.com/@manurubo/sunburst-con-zoom. El chord diagram se

puede encontrar en solitario con el código en: https://observablehq.com/@manurubo/chord-diagram.

Se han seguido las interactividades definidas en el mockup, con dos añadidos:

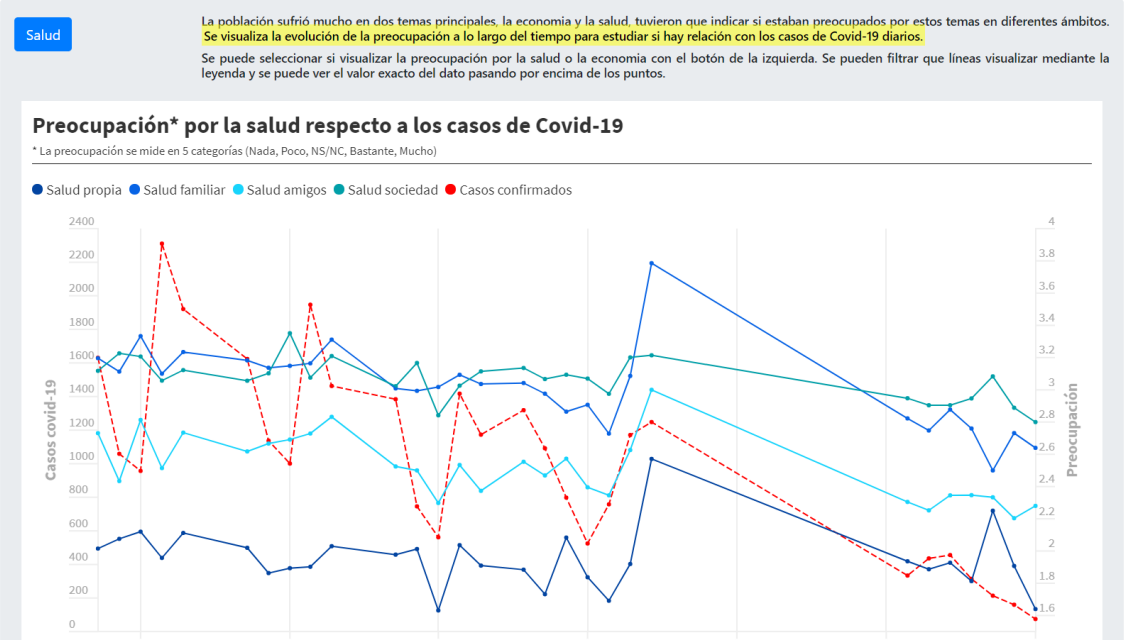
- Obviamente cada link del diagrama de cuerdas y cada segmento del sunburst permiten interactividad.
- **Con los botones para definir las jerarquías del sunburst, se permite que la segunda jerarquía esté en blanco, con lo que se presenta directamente los sentimientos principales para cada categoría de la primera jerarquía.**

La tercera visualización se ha realizado mediante Flourish y se puede encontrar en solitario en: https://public.flourish.studio/visualisation/6194291/. A esta visualización se ha mantenido la interactividad diseñada, pero con dos modificaciones:

- Evidentemente, ahora cada punto tiene la posibilidad de mostrar detalles bajo demanda, al contrario que en el mockup que solo eran dos como ejemplo.
- **Se añade una leyenda a la visualización que permite reconocer que representa cada una de las líneas pintadas,** de manera que los casos confirmados tienen un color rojo. Pero todas aquellas variables que hablan de la preocupación de economía en diferentes ámbitos, comparten diferentes tonos de azul, para que se pueda identificar rápidamente que todas esas líneas hablan de preocupación de la economía aunque en diferentes ámbitos. **Esta leyenda es interactiva, de manera que seleccionar un elemento en la leyenda permite mostrar o no esta línea.**

Como extra de interactividad para esta gráfica, se reconoció que la preocupación por la salud (p_31_(1-4)) no se había contemplado a pesar de que tenía la misma estructura de preguntas que la preocupación por la economía. **Por tanto, se generó la misma gráfica pero para preocupación por la salud (https://public.flourish.studio/visualisation/6195588/), y mediante Javascript y Jquery, se muestra la gráfica de una preocupación o otra en función de un boton de Bootstrap.**

Y si clickamos en el botón:



Vemos como ha cambiado el valor del botón y la gráfica, que ahora contiene la preocupación por salud.

En definitiva, y si se interactúa con la visualización de datos, se puede apreciar que las diferentes visualizaciones realizadas permiten manipular los elementos interactivos de las visualizaciones. Permitiendo el mantra de "Overview first, zoom and filter, and details on demand." y siguiendo una serie de buenas prácticas definidas en la asignatura como el uso reducido de líneas en una visualización que solo aportan ruido, o la utilización de los colores para agrupar elementos similares. Además, esto se ha realizado mediante 2 herramientas de las proporcionadas (Observable y Flourish), e incluso se ha añadido una parte de interactividad más, ya que algo que no permite Flourish, como es el seleccionar datos de la gráfica por interactividad, se ha implementado mediante librerias de Javascript.

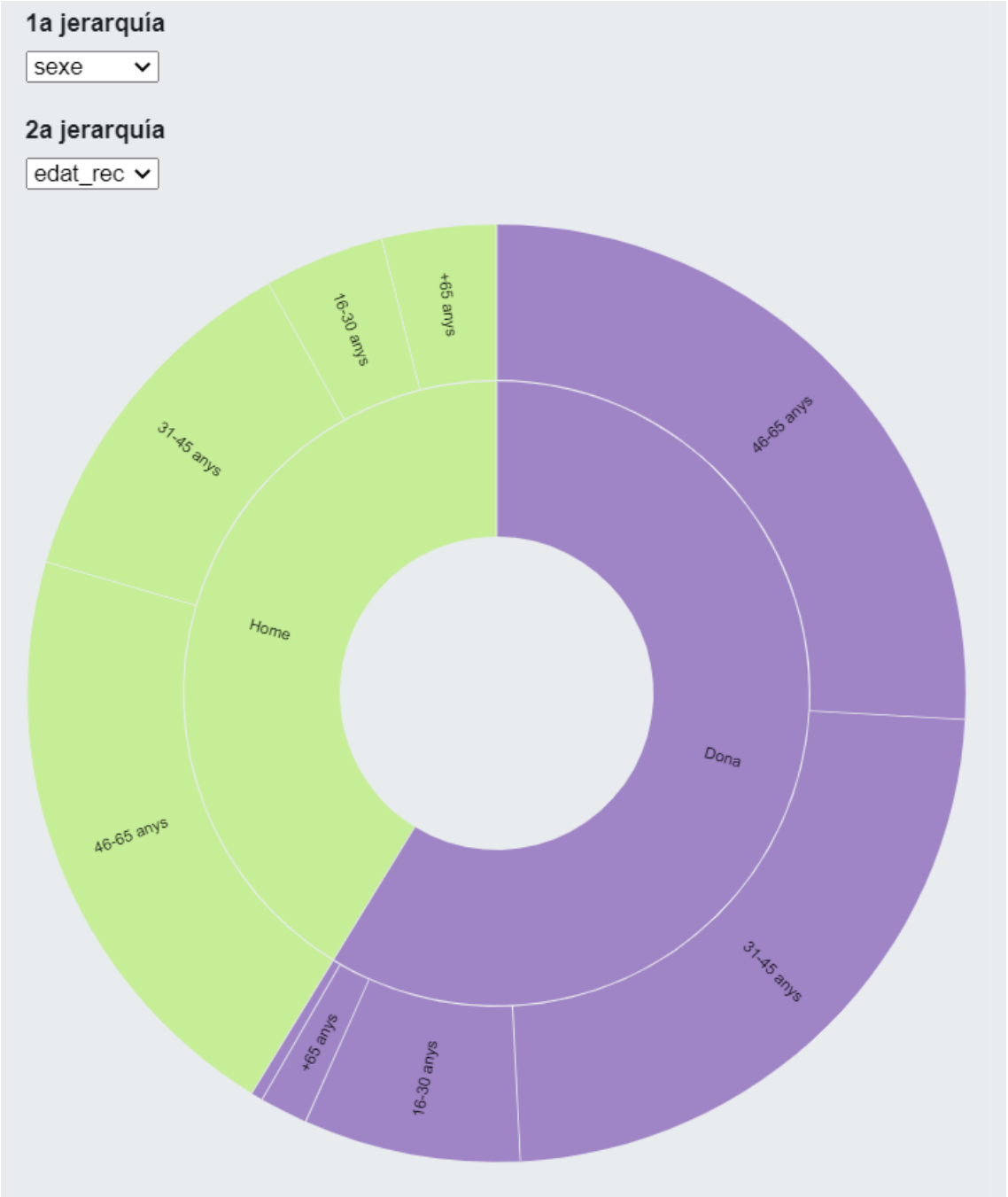
4. Funcionalidades de la visualización

La visualización de datos consta de 3 visualizaciones, por lo que se va a plantear 3 preguntas diferentes para el dataset y se va a mostrar como resolverlas cada una con una visualización.

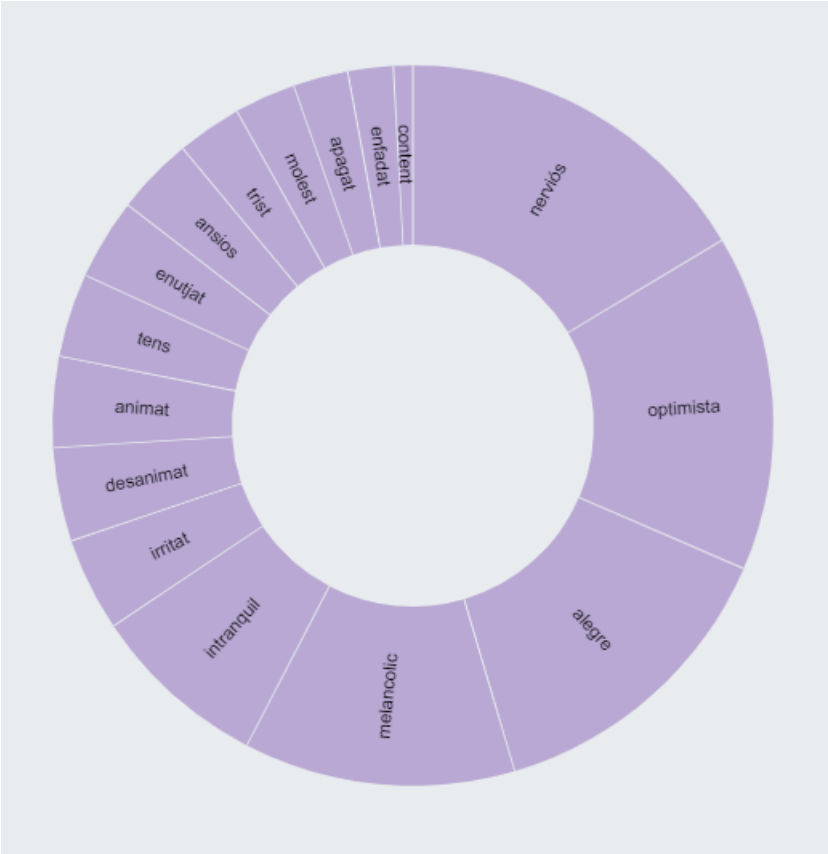
1. **¿Cuáles fueron los 3 sentimientos principales que sentían las mujeres de entre 46 y 65 años durante el confinamiento en Cataluña?**

Esta pregunta es interesante en el conjunto de datos, porque el dataset contenía información sobre que puntuación de 16 sentimientos diferentes sentía la persona. Se ha entendido que el sentimiento con mayor puntuación es el sentimiento principal que sentían los encuestados. Para comprender como se podía sentir una parte de la población, podemos definir unas jerarquías y ver los sentimientos principales de el grupo.

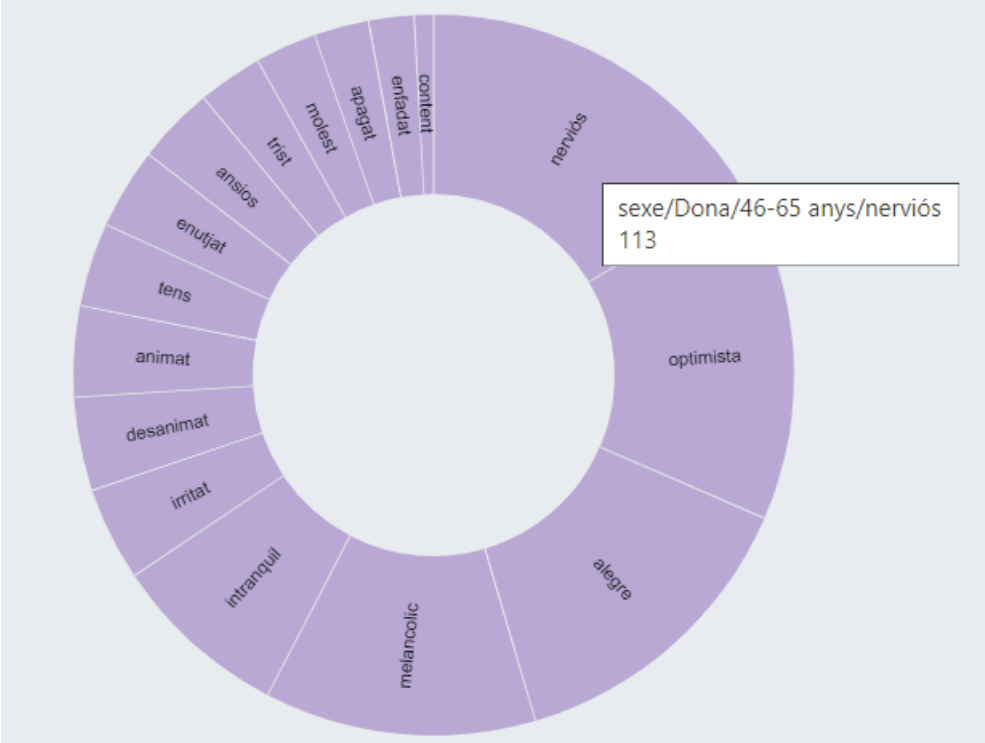
En este caso, primero se debería de seleccionar en los botones del Sunburst, como primera jerarquía "Sexe" y como segunda "edat_rec". De esta manera, la visualización muestra ambas jerarquías.



Ahora, se tiene que seleccionar directamente el segmento morado (Dona) de 46-65 años, que representaría mujeres de esa edad. Aunque se puede hacer por partes, primero seleccionando Dona, y en la siguiente interacción la franja de edad que interesa. De esta manera, ya tenemos la respuesta a nuestra pregunta, y vemos los sentimientos principales para estas mujeres de 46 a 65 años durante el confinamiento:



Vemos que los 3 sentimientos principales de estas mujeres eran nerviosa, optimista y alegre. Además, la visualización nos permite situarnos sobre un segmento y ver cuantas mujeres de esas edad tenían un sentimiento como mayoritario, por ejemplo, poniéndonos encima de la fracción de nervios:



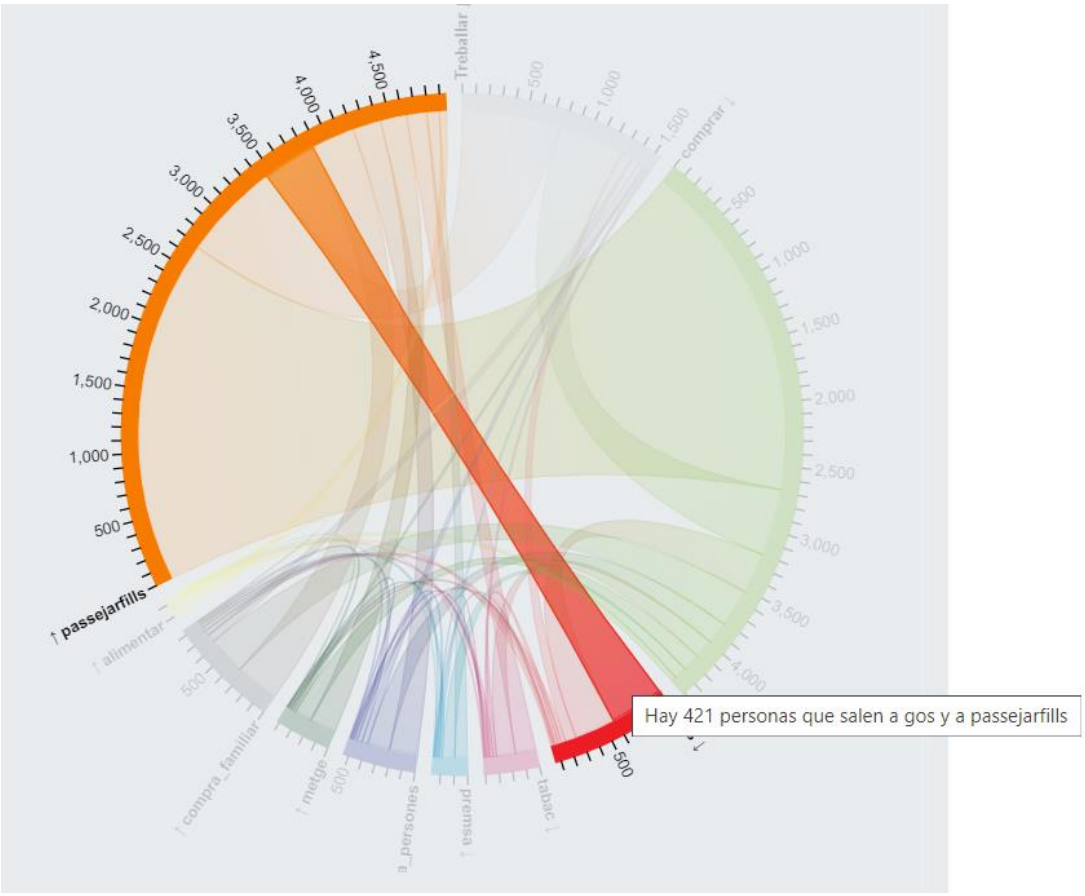
Se ve que hay 113 mujeres de esa edad que estaban principalmente nerviosas durante el confinamiento, siendo este su sentimiento principal.

2. **¿Cuántas personas que salen del confinamiento para pasear al perro salen también a pasear a sus hijos?**

Esta pregunta puede ser interesante de resolver porque nos permite identificar si había relaciones entre las diferentes salidas de las personas durante el confinamiento. Por ejemplo, encontrar que mucha gente que salía por una razón, salía mucho por otra razón o no.

Esta pregunta se encuentra en el dataset sobre todo en las preguntas p_25_$razon. Donde las personas indicaban si habían salido por el motivo o no para varias razones.

Para solucionar la pregunta, simplemente tenemos que ponernos sobre el link en el diagrama de cuerdas, y se disminuye la visualización del resto de links, mientras que solo se muestra este, los grupos a los que pertenece y el número de personas que comparten estas dos salidas.



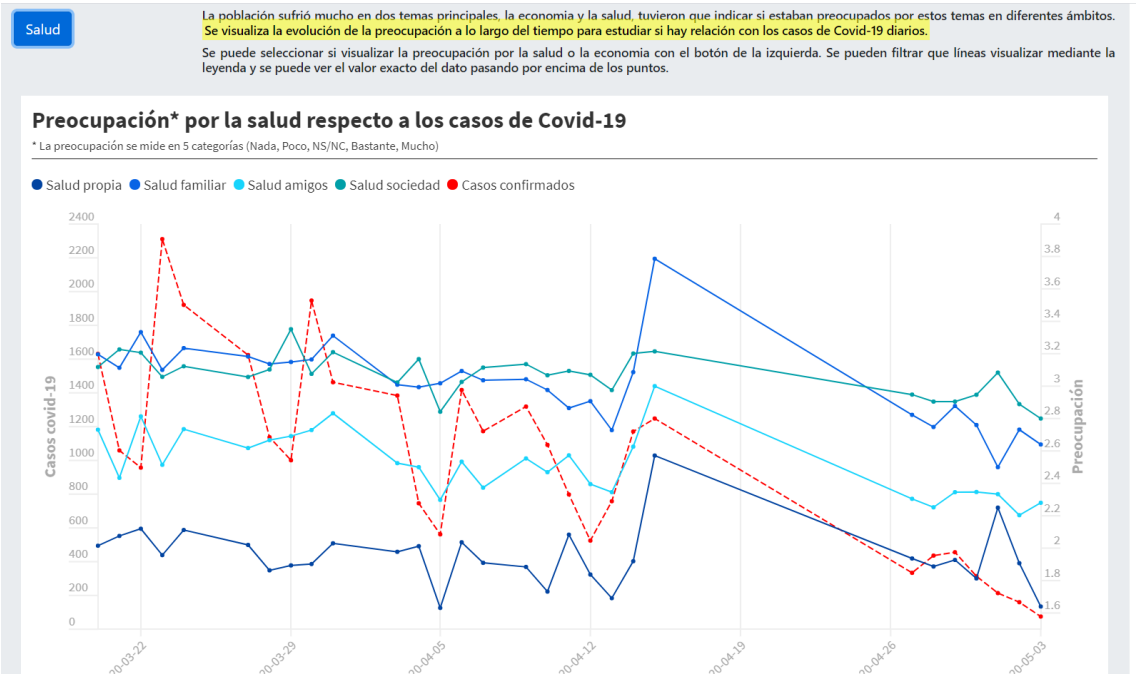
Mientras que la respuesta a esta pregunta es muy simple, hay 421 personas que comparten estas salidas, la visualización simplemente con esto nos permite identificar muchas más cosas. Es interesante ver como mientras que esta relación no es la mayor de todas (la relación entre pasear hijos y comprar es enorme), es una relación que se ha dado bastante siendo la 4ª más grande de todas las relaciones. Además, vemos cómo supone que casi toda la mitad de personas que salieron por pasear un perro también salieron a pasear a los hijos, pero para pasear a los hijos es solo su tercera relación mayor, suponiendo estas 421 personas un menos de un 10% de todas sus relaciones.

3. **¿Hay relación visual entre la preocupación por la salud familiar de los encuestados y los casos de Covid-19?**

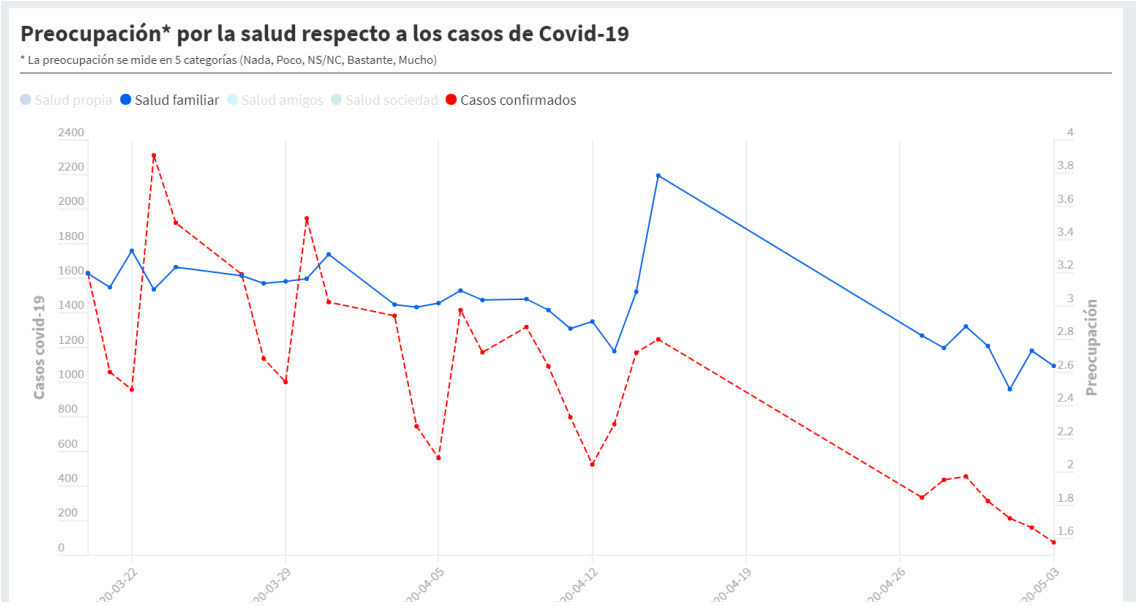
Esta pregunta es interesante de responder porque obviamente, es lógico pensar que las personas se preocuparan más cuando había más casos de Covid. El dato de preocupación por salud lo tenemos en el dataset 1 mediante las p_31, y el dato de casos de Covid, lo tenemos en el dataset 2.

Mediante la visualización 3, de líneas temporales podemos responder a esta pregunta. Inicialmente la gráfica muestra la preocupación por la economía:

Se ha de pulsar sobre el botón para visualizar la gráfica para la preocupación de salud. Esta gráfica ya nos permite ver todas las preocupaciones de salud respecto a los casos de covid:



Si queremos visualizar solo la preocupación por la salud familiar y los casos confirmados, tenemos que seleccionar en la leyenda el resto de elementos, para que desaparezcan sus líneas temporales.



Ahora, ya solo vemos la salud familiar y los casos confirmados. Vemos que, en las primeras semanas, ambos disminuyeron poco a poco, tanto los casos como la preocupación por la salud familiar. A partir del 2 de mayo, los casos comienzan a disminuir mucho más que lo que disminuye la preocupación por la salud familiar. También vemos que el 13, 14 y 15 de Abril hubo una subida de casos cuando parecía que comenzaba a disminuir el número de casos, lo que hizo que los días 14 y 15 de Abril la preocupación por la salud familiar ascendiera mucho. Por último, vemos como ya en los últimos puntos, tanto los casos como la preocupación ha disminuido a sus números más pequeños.

Por tanto, se puede concluir que hay cierta relación, ya que generalmente ambas líneas descienden a la par, y que cuando los casos subieron tras una bajada, la población catalana se preocupó más.