VISUALIZACIÓN DE DATOS

PRA2 – Creación de la visualización y entrega del proyecto

Diego Martín Montoro

1. Título de la visualización

Proceso** de **vacunación** mundial **COVID-19.

2. URL de la visualización y el código

Enlace de la visualización:
https://public.tableau.com/profile/diego8806#!/vizhome/Practica2_16217884801600/Dashboard1?publish=yes

Enlace git con todos los ficheros: https://github.com/Diego-Martin-Montoro/Visualizaci-n-PRA2

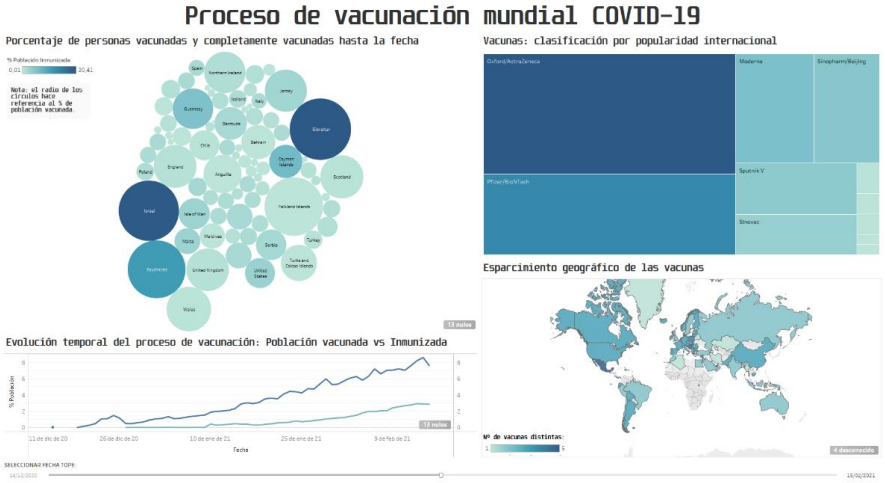
3. Descripción corta del documento y de lo que se presenta

En este documento, se contestan a las cuestiones solicitadas en el enunciado de la práctica. Además, se ilustra cómo se ha procedido en el proceso de selección de las técnicas de visualización que compondrán el producto final.

Para ello, se acompaña la entrega con un notebook en html en el cual se puede presenciar el proceso de pre-procesado de datos de manera razonada y el código que lo sustenta, para acabar con la exportación de los datos de interés a dos archivos .csv que alimentarán las distintas visualizaciones que se han generado.

4. Explicación de la visualización. ¿Qué visualización se ha escogido y por qué?

La visualización escogida es la siguiente:



Como podemos ver, esta se compone de cuatro gráficas dinámicas. La justificación de la elección de cada de ellas se basa en los principios de percepción humana y en el deseo de maximizar la comodidad y la efectividad de la visualización desarrollada. Estas razones son las siguientes:

- Gráfica de esferas: el tamaño es algo con lo que el ser humano está acostumbrado a tratar, y rápidamente este es capaz de discernir qué objetos tienen un tamaño mayor. Como uno de los factores que nos interesa estudiar es la cantidad de vacunados que hay, representar esta medida como el tamaño de una esfera para cada país permite reconocer rápidamente qué países van a la cabeza.

- Mapa en árbol: por la misma razón que para la gráfica de esferas, y al ser nuestro deseo el de mostrar el ranking de vacunas por popularidad centrándonos en las más populares, este tipo de gráfico encaja a la perfección. Con un simple vistazo vemos qué vacunas son líderes a nivel mundial.

- Gráfica de líneas con dimensión temporal: debido al interés en visualizar la evolución temporal de las cantidades de vacunas administradas y de los porcentajes de inmunización, un gráfico de líneas con eje horizontal temporal es una opción que encaja a la perfección. En él, rápidamente podemos observar cómo fluctúan los valores de estas medidas en el tiempo, siendo un gráfico sencillo de interpretar y potente.

- Mapa de calor: debido al interés en estudiar el progreso de vacunación de manera geográfica, con especial hincapié en la distribución de las distintas vacunas, usar una representación geográfica de los datos era una elección adecuada. Gracias al mapa podemos ver como se distribuyen geográficamente las distintas vacunas, y pueden observarse ciertos grupos de naciones con claras preferencias de origen político.

En todas las gráficas la intensidad de color se usa como un apoyo a la percepción del usuario. Indicando esta la magnitud de las medidas visualizadas. Todos estos elementos usados en conjunto permiten realizar un análisis de los procesos de vacunación a nivel mundial a la vez que se estudia el uso de las distintas vacunas sin apenas esfuerzo para el usuario.

Había otras opciones como por ejemplo usar diagramas de barras en vez del mapa en árbol o la gráfica de esferas, pero considero que estos son mucho más atractivos incitando así a los usuarios a interactuar con la visualización. Esta elección dificulta el establecimiento de un ranking como tal (sobre todo en el caso de las esferas) pero aumenta mucho la atractividad de la visualización además de que establecer rankings no es nuestro objetivo. Nuestro objetivo es detectar valores muy grandes o pequeños y eso se puede hacer perfectamente.

5. Explicación de la visualización. ¿Para qué y para quien puede servir esta visualización?

Esta visualización puede resultar muy útil para los organismos internacionales de la salud (p.e OMS). Concretamente, con esta visualización podrían monitorizar el proceso de vacunación con el objetivo de detectar ciertas situaciones:

1. Países poco eficientes en el progreso de vacunación. Países que, tienen un índice de vacunación bajo, o que, aun teniendo un índice de vacunación alto, tengan una anómala proporción de personas inmunizadas (más baja de lo que debería). La segunda dosis hay que administrarla pasado un tiempo determinado, si dicha dosis se demora es posible que el efecto no sea el deseado.

2. Detectar países muy eficientes en el proceso de vacunación: podría estudiarse el proceder de estos para ayudar a otros países menos eficientes a mejorar en su desempeño contra el virus.

3. Países poco colaboradores. Países que no aportan datos internos de su sistema de salud, o cuyos datos aportados presentan comportamientos extraños o poco usuales. P.E: China no aporta cifras reales, y en la mayoría de las ocasiones no aporta los porcentajes de vacunados y completamente vacunados.

4. Detectar países poco previsivos y/o responsables. Usar muchos tipos de vacuna distintos en un país puede ser una señal de falta de previsión. México usa 5 vacunas distintas, y algunas de ellas de dudosa fiabilidad.

5. Ver qué países usan un tipo de vacuna concreto: facilitaría la monitorización de estas a nivel geográfico. También permite extrapolar el estudio al contexto político. P.E: vemos como los países “aliados” de Rusia usan la vacuna Sputnik V y no las europeas o americana.

6. Detectar brechas sociales: zonas del primer mundo con unos procesos de vacunación muy avanzados y zonas del tercer mundo (o del segundo) con graves déficits en el proceso, o incluso con un proceso de vacunación inexistente. Gracias a esto podrían comenzarse planes internacionales de rescate y apoyo a dichas localizaciones.

6. Explicación de la visualización. ¿Qué preguntas responde la visualización?

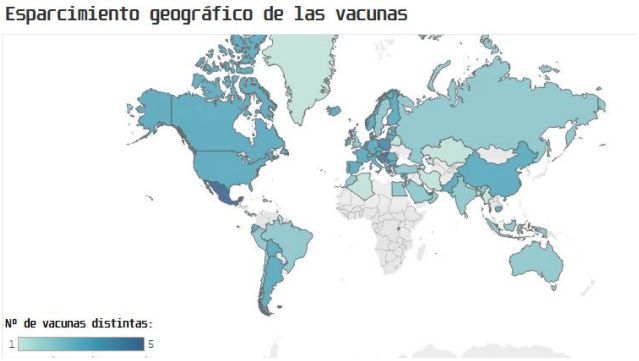
- ¿Qué vacuna es la más popular?, ¿Y la menos popular?
  La más popular es Oxford/Astrazeneca, la menos es EpiVacCorona (usada en Rusia).

- ¿Cuál es el país más eficiente vacunando de América?, ¿Y de Europa?
  Estados Unidos con un 25.78% de la población inmunizada el día 20 del mes pasado.
  De Europa es Mónaco, con un 27.67% de la población inmunizada el día 20 del mes pasado.

- ¿Se observan preferencias geográficas por alguna vacuna en especial?
  Sí. La vacuna Moderna, usada mayoritariamente en países del 1er mundo, principalmente en Europa.

- ¿Hay una brecha significativa entre el primer mundo y el tercer mundo en lo que a vacunaciones se refiere?
  Sin duda. Hasta el día 22 de enero no hubo ni una sola vacuna en África mientras que otros continentes como Europa o América ya habían realizado casi en su completitud (geográfica) vacunaciones a lo largo de diciembre de 2020 y los primeros días de enero del 2021.

  Adjunto fotografía para el día 15 de febrero:



  Se puede apreciar el gran vacío africano en vacunaciones.

7. Descripción técnica del proyecto: ¿Qué transformación de datos ha habido que hacer respecto al juego de datos inicial?

Para realizar las gráficas de evolución temporal y esferas no ha sido necesario preprocesar los datos de ningún modo. Sin embargo, para las otras dos gráficas ha sido conveniente un cambio en la representación de estos.

En concreto, he desenrollado la variable 'vaccines'. En el formato original, esta variable aparece como una lista de nombres de vacunas separadas con el carácter ','. El desenrollado ha consistido en generar una fila por cada nombre de vacuna, replicando el resto de los datos del registro tantas veces como vacunas haya. Para el caso que nos ocupa, los únicos datos que se han conservado son: el país, la vacuna y la fecha del primer registro de dicho país y dicha vacuna (de manera que podemos saber cuándo un país comenzó a usar una vacuna determinada).

En el notebook adjuntado en la entrega (*preproceso.html*) se puede ver el proceso.

8. Descripción técnica del proyecto: ¿Qué lenguaje, librería, software has usado y por qué?

El software que he usado para el preprocesamiento de datos es Python, en concreto los módulos numpy y pandas con los que he realizado las transformaciones comentadas en el punto anterior. Considero que es el mejor lenguaje/librerías en la actualidad para el preprocesamiento de datos ya que son herramientas potentes, fáciles de usar y con una gran comunidad.

Para la creación del dashboard a partir de los datos preprocesados en Python he usado el software Tableau porque además de parecerme una opción muy buena debido a su alta variabilidad y a la gran agilidad que aporta al realizar visualizaciones, como comenté en el foro, me es de gran interés laboral el profundizar en el uso de este.

En esta ocasión he profundizado en el funcionamiento de este, llegando a programar de manera "manual" los filtros, consiguiendo una interacción con el usuario más fluida. La forma tradicional o más directa de realizar filtros limita mucho las posibilidades de filtrado y proyección de datos. Una de las mayores limitaciones es que los filtros tradicionales solo pueden aplicarse (correctamente) a gráficas que se nutran de datos con formatos idénticos, por el contrario, usando parámetros calculados podemos filtrar lo que nos haga falta sin preocuparnos del formato de los datos. Además, el uso de estos filtros básicos fuerza al usuario a confirmar la mayoría de las acciones limitando así en gran medida la fluidez de trabajo.

9. Descripción técnica del proyecto: ¿Qué proceso he seguido para desarrollar la visualización?, ¿Qué nuevas técnicas he usado para mejorar mis visualizaciones?

Es mi voluntad presentar un pequeño resumen que describa cómo ha sido el proceso de desarrollo de esta visualización ya que he probado una praxis distinta en varios aspectos.

El proceso comienza con una exploración rápida de los datos y las visualizaciones que pudieran aportar mayor capacidad de extracción de conocimiento. Acto seguido, y una vez elegidas las visualizaciones a realizar, estudié qué preprocesamientos eran pertinentes aplicar a los datos y los desarrollé en un notebook de Python.

Luego, volqué los datos preprocesados en el libro definitivo de tableau. De manera gradual, fui realizando las visualizaciones escogidas. Poco a poco, y como capas en una cebolla, fui moldeando estas añadiendo features o eliminanto otras, de manera en que maximizase tres conceptos: productividad, atractividad y comodidad. De esta manera, genero la visualización usando un enfoque *Agile*, donde yo hago de *product owner* y de *customer*. Aplicando cambios o ampliaciones, realizando pruebas, aceptando aquello que mejora el producto y eliminando o corrigiendo lo que lo empeora.

En los siguientes puntos, explico algo más en detalle puntos de cierta relevancia en el proceso expuesto:

A. Proceso de exploración más ágil

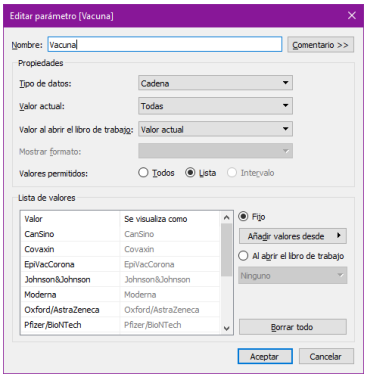
En las otras asignaciones, me percaté de que en ocasiones me trababa demasiado tiempo con errores tontos o corrigiendo datos (formatos, errores) realizando la exploración de datos en lenguajes de programación como puede ser Python. Esta vez la exploración ha consistido en cargar los datos en un libro de tableau para generar muchas visualizaciones muy simples, aprovechando la alta tolerancia a errores en los datos de esta herramienta para agilizar el proceso.

De este modo, en un intervalo de tiempo muy corto (no llega a media hora) evalúo diferentes “propuestas” de visualizaciones para los datos sin necesidad de entrar a detalles ni de pelearme con errores o bibliotecas. Rápidamente y gracias a la familiaridad obtenida en la PRA1 con respecto al conjunto de datos, supe qué gráficos eran los más adecuados para la visualización final.

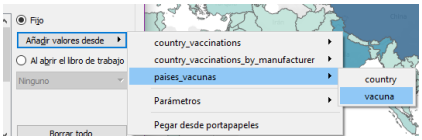
B. Realización manual de filtros usando parámetros

Como indiqué en el apartado anterior, programar de manera “manual” los filtros se traduce en una interacción de la visualización con el usuario más fluida. Además de esta manera ahora se puede aplicar un mismo filtro a visualizaciones que se nutran de distintas fuentes de datos (con distintos formatos), aumentando la agilidad de desarrollo (habrá que gestionar menos filtros) y también la potencia del sistema de filtros de la visualización.

Voy a ilustrar cómo he configurado estos filtros, poniendo como ejemplo el proceso de construcción del filtro para las vacunas.
Primero se crea el parámetro Vacuna, con la siguiente configuración:

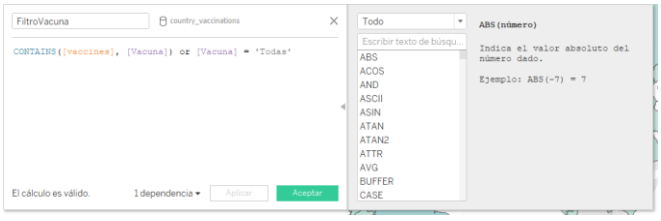


La lista de valores la he extraído del campo vacunas del dataset Preprocesado:



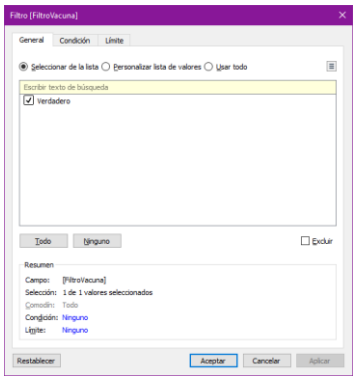
De manera manual, hay que añadir el valor “Todas”.

Una vez creado el parámetro, debemos crear un campo calculado con la siguiente definición:



Si nos fijamos en el código, vemos que lo que hace es comprobar si el valor de parámetro vacuna está contenido en la cadena ‘vaccines’ (para cada dato), o si el valor del parámetro es ‘Todas’.

Ahora, solo hace falta añadir este campo calculado como filtro:



Seleccionando la opción ‘verdadero’, haremos que solo se muestren los datos cuyo campo calculado valga verdadero (en el caso de que haya seleccionada una vacuna).

Lo bueno de los parámetros es que son generales, como variables globales en un código. De manera que pueden ser consultados o modificados desde cualquier sitio. Además, el hecho de usar un campo calculado para la realización del filtro nos brinda toda la potencia del uso de código para así generar la lógica que deseemos prácticamente sin limitaciones.

Gracias a esto he podido desarrollar un filtro conjunto de vacunas y países, usando operadores de igualdad y de contenencia de subcadenas.

Para finalizar, la visualización va muchísimo más fluida usando este método: se ve que procesa rápidamente las operaciones lógicas aplicadas en campos calculados ya que con los filtros “tradicionales” todo iba más forzado.

Gracias a este nuevo enfoque he podido extraer más conocimiento del conjunto de datos bajo observación. Gracias a estas nuevas herramientas, he podido prescindir del conjunto de datos más pequeño (el que se basaba únicamente en las vacunas y tenía información limitada a un número reducido de países) y extraer esta información del conjunto de datos principal obteniendo unos datos más fieles a la realidad, de muchos más países y con una componente temporal con una granularidad más fina.

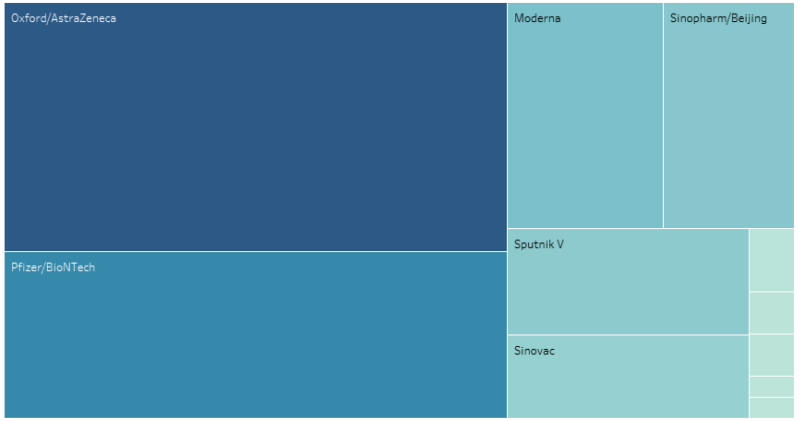
10. Presentación de la visualización de datos.

En las páginas siguientes voy a describir la visualización comentando los distintos puntos de interés, además adjuntaré capturas de pantalla para apoyar la explicación.

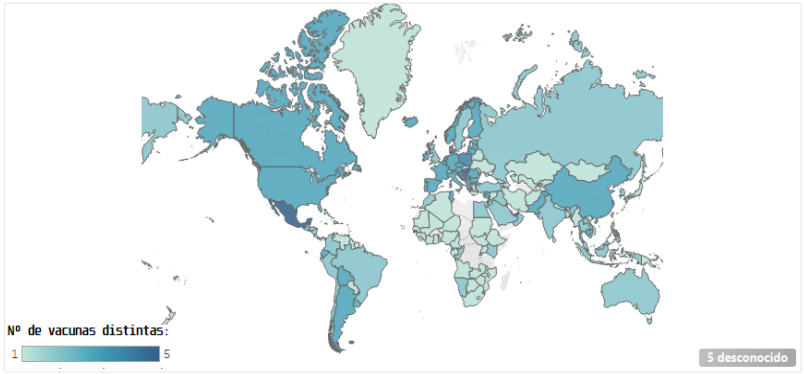
Proceso de vacunación mundial COVID-19

Porcentaje de personas vacunadas e inmunizadas hasta la fecha (Todas)

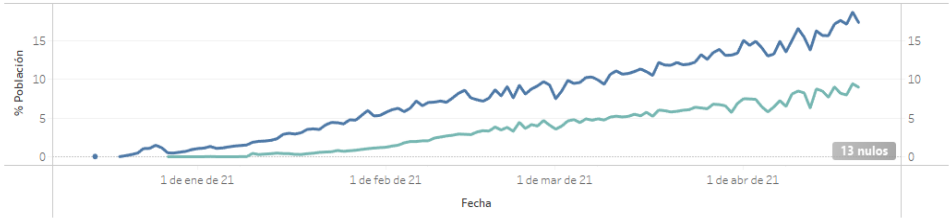
Vacunas: clasificación por popularidad internacional



Esparcimiento geográfico de las vacunas (Todas)



Proceso de vacunación: Población vacunada vs Inmunizada (Todas)

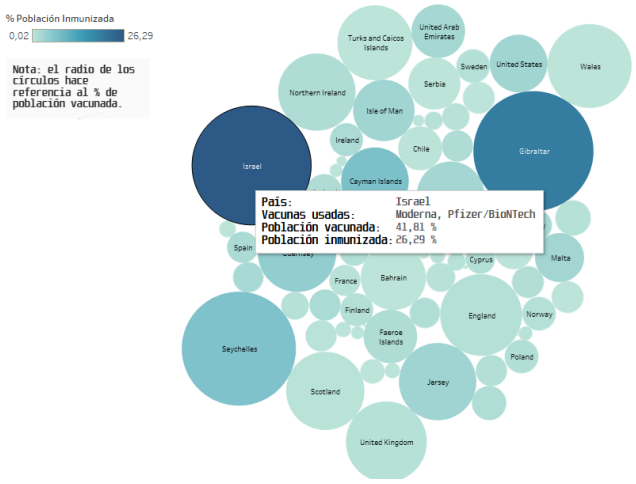


SELECCIONAR FECHA TOPE:

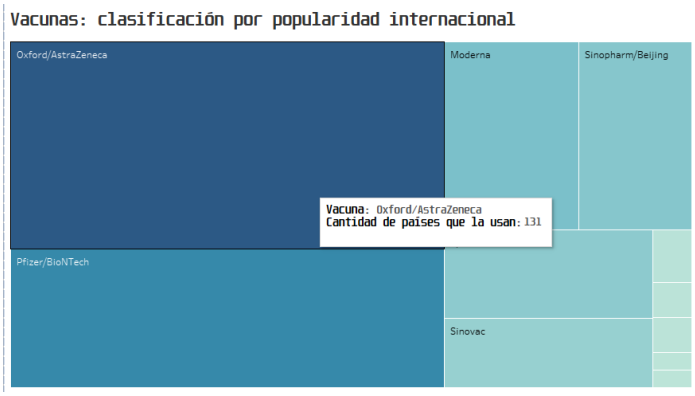
14/12/2020 — 20/04/2021

Capturas:

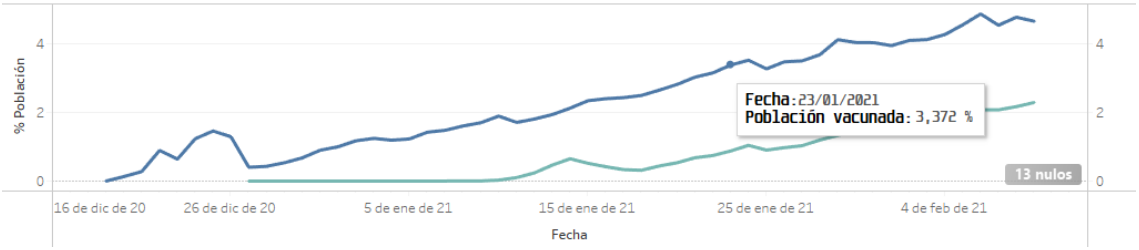
Porcentaje de personas vacunadas e inmunizadas hasta la fecha (Todas)



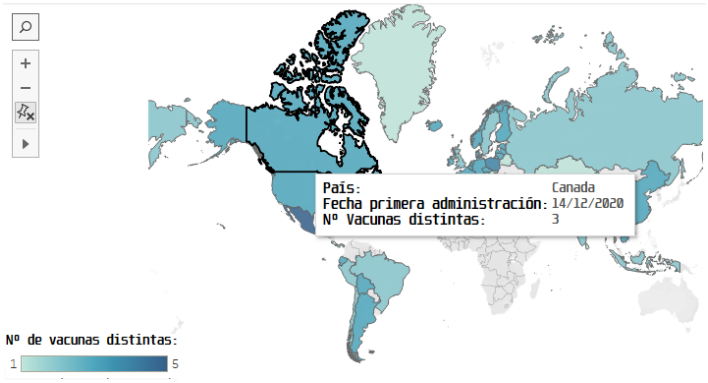
Vacunas: clasificación por popularidad internacional



Proceso de vacunación: Población vacunada vs Inmunizada (Moderna)



Esparcimiento geográfico de las vacunas (Todas)



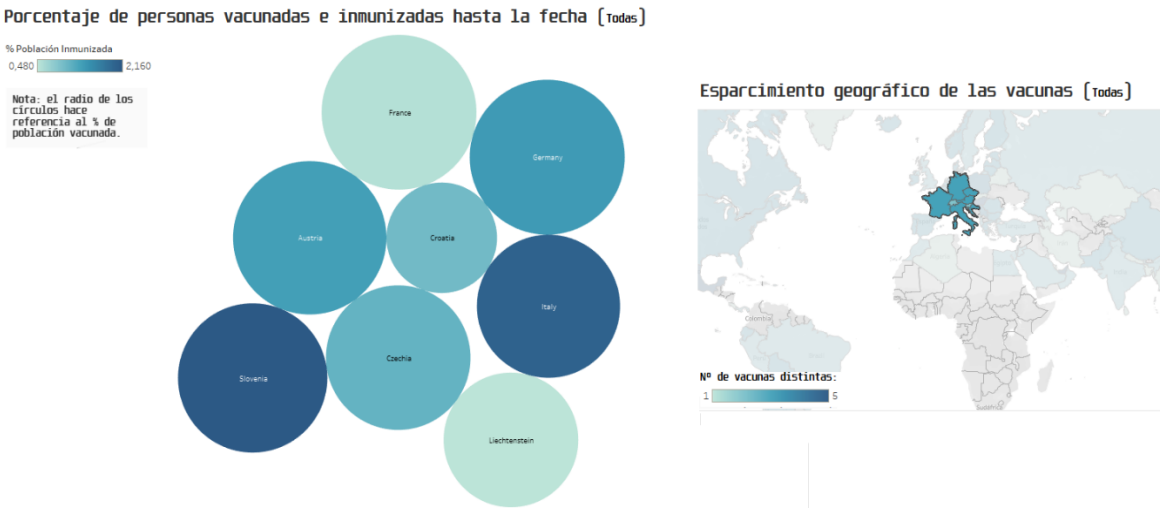
Como podemos ver el dashboard se compone de 4 visualizaciones distintas:

1. **Mapa geográfico:** en él observamos geográficamente como se han distribuido las distintas vacunas existentes hasta la fecha tope indicada en el filtro temporal. En caso de estar visualizando los datos para todas las vacunas, nos permite reconocer rápidamente mediante la intensidad de color que naciones están usando una mayor variedad de vacunas contra el COVID.

   Una vez **situado el ratón sobre una región**, se aporta información extra con un bocadillo emergente: cantidad de vacunas distintas usadas en dicha región y fecha de comienzo del proceso de vacunación.

   El gráfico nos permite realizar zoom o desplazarnos para centrar la atención donde deseemos.

   Si seleccionamos una región o regiones (pinchar y arrastrar) entonces la gráfica de esferas y la de evolución temporal del proceso de vacunación filtrarán sus datos a únicamente aquellos que correspondan con los países seleccionados. Ejemplo:

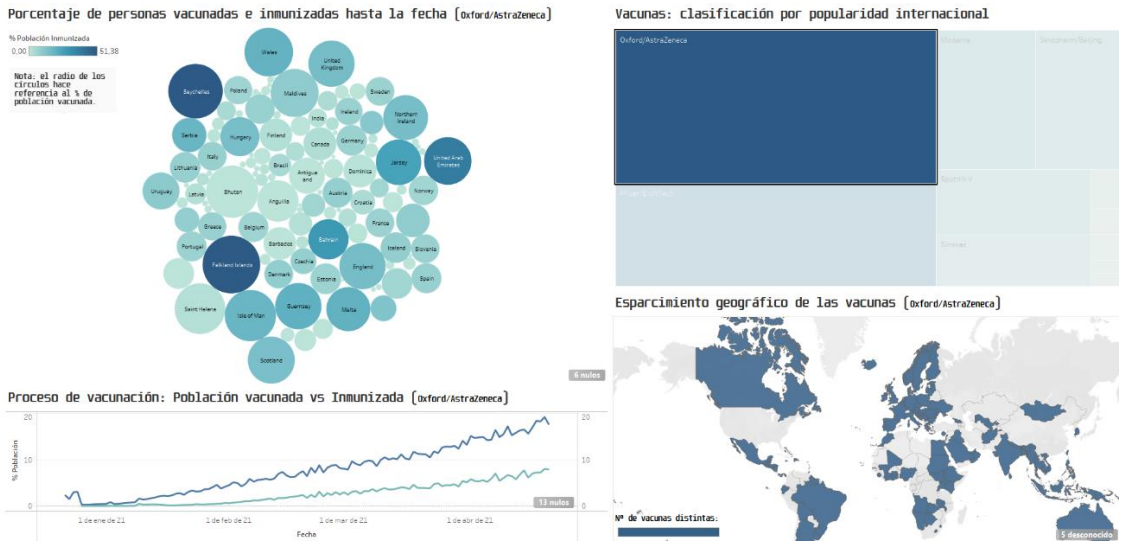


   Para deseleccionar basta con clicar fuera de las regiones seleccionadas.

2. **Mapa de árbol:** permite observar rápidamente qué vacunas son o han sido más populares. Tanto el tamaño como la intensidad del color de las cajas reflejan el volumen de países que han usado una vacuna determinada hasta la fecha tope indicada en el filtro temporal.

   Una vez **situado el ratón sobre una celda**, se mostrará información extra: cantidad de países que han usado dicha vacuna (numéricamente), de modo que permite concretar cantidades.

   Además de elemento de visualización, **también sirve como selector o filtro**, ya que, seleccionando una vacuna determinada automáticamente el resto de las visualizaciones solo mostrarán datos correspondientes a dicha vacuna:



Para deseleccionar basta con volver a clicar en la caja, o clicar al aire.

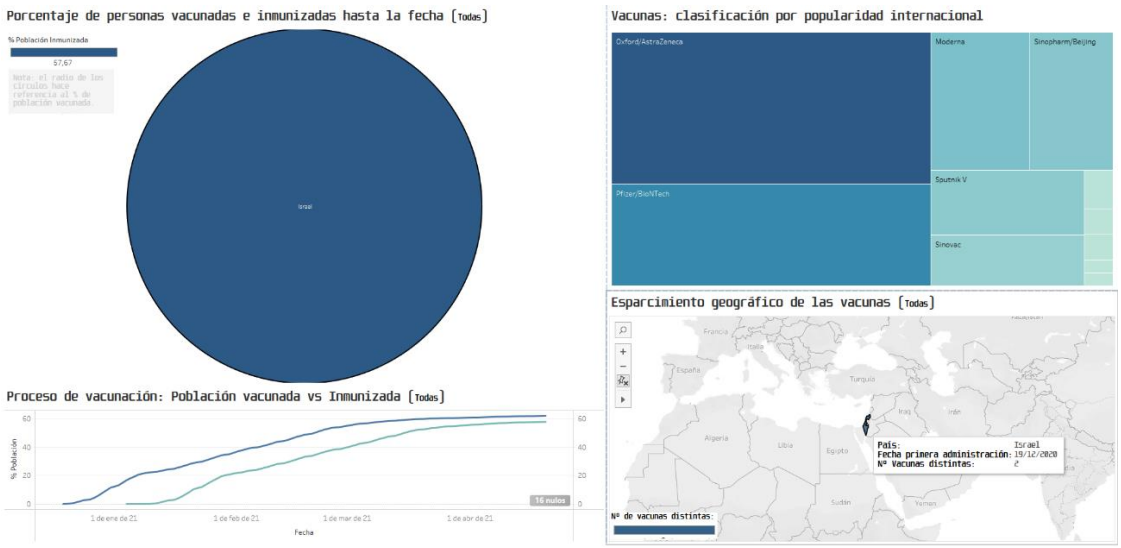
NOTA IMPORTANTE: cuando filtramos por una vacuna determinada, lo que hacemos es quedarnos solo con aquellos países que la usen (entre otras), no aquellos que **únicamente** usen dicha vacuna.

3. **Gráfico de esferas:** permite observar qué países se encuentran en un estado más avanzado en el proceso de vacunación en la fecha límite establecida en el filtro temporal. El tamaño de las esferas hace referencia a la proporción de personas vacunadas en el país, mientras que la intensidad de color se enfoca en mostrar el valor de la proporción de personas inmunizadas (hasta la fecha).

   Una vez **situado el ratón sobre una esfera**, se mostrará información extra: proporción de personas vacunadas e inmunizadas en el país (numéricamente), de modo que permite concretar cantidades.

   Entonces, esferas grandes y muy oscuras indicarán países o regiones que van muy por delante de todas las demás en el proceso de vacunación. Por ejemplo: Gibraltar, o Israel. Análogamente, **esferas muy grandes y claras reflejarán un claro déficit en el proceso de administración de la segunda dosis.** Ejemplo: Santa Elena o Bhutan.

   Si se selecciona una esfera (país) automáticamente las gráficas inferiores pasarán a mostrar solo la información de dicho país, es decir: su evolución temporal del proceso de vacunación, y su localización y cantidad de vacunas distintas usadas en el mapa:



Para deseleccionar basta con volver a clicar la esfera.

4. **Gráfico de líneas:** permite ver cómo evolucionan los porcentajes de vacunados e inmunizados a través del tiempo, hasta la fecha límite establecida en el filtro temporal.

   Si solo estamos visualizando los datos para un país, entonces tendremos sus porcentajes vírgenes. Sin embargo, si estamos visualizando datos generales, los porcentajes serán los promedios mundiales. De manera análoga ocurre con la vacuna seleccionada (una determinada o todas).

   Una vez **situado el ratón sobre un punto de una línea**, aparecerá información extra en un bocadillo: fecha concreta y valor de la métrica pertinente (vacunados o inmunizados).