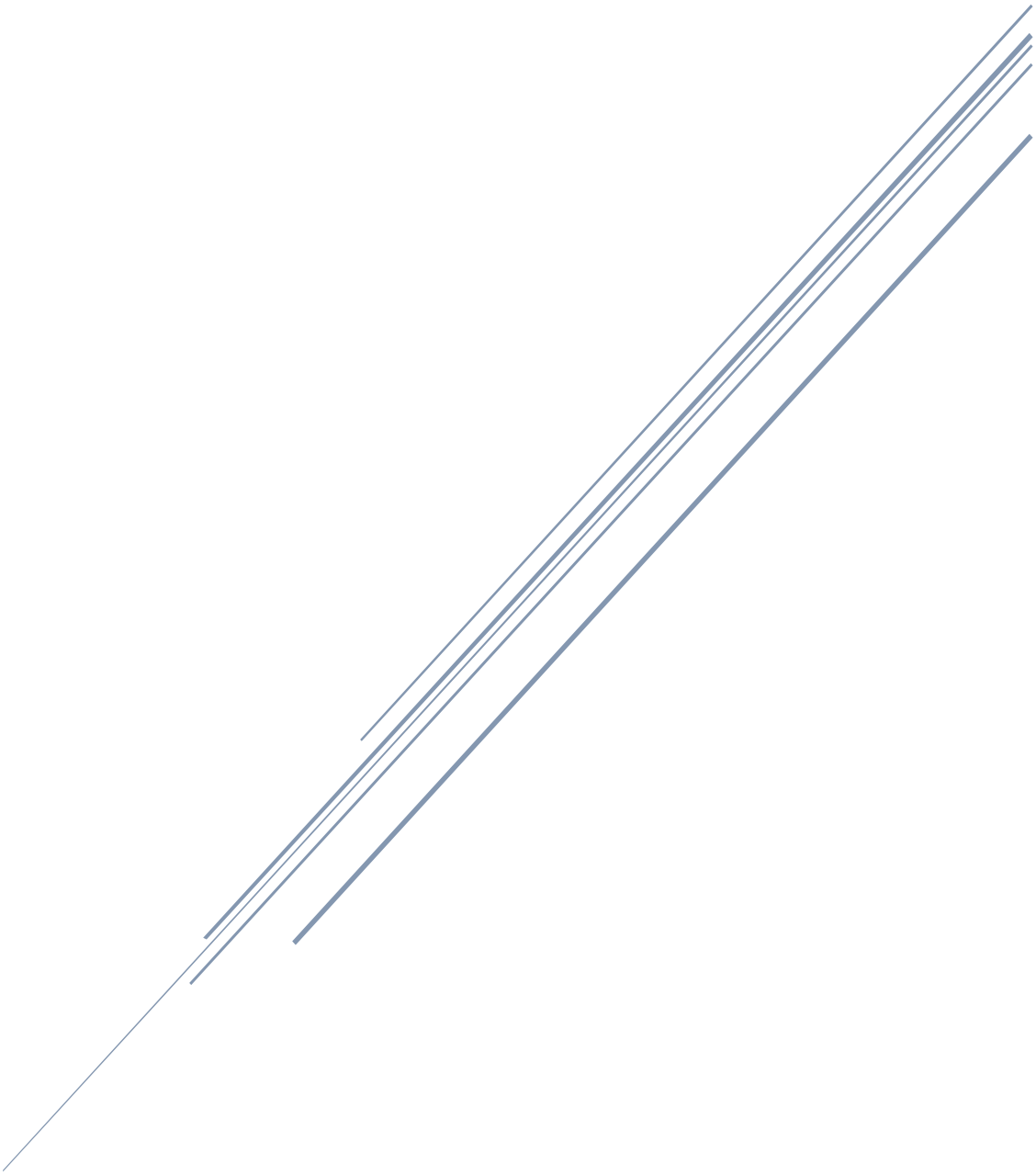
VISUALIZACIÓN DE DATOS

PRA2 – Dashboards “COVID-19 World Vaccine Adverse Reactions”



UOC – MSc Data Science
Álvaro Ortiz Hernández

PRESENTACION

Título de la visualización

Efectos adversos de la vacuna del Covid-19 en EEUU.

URL de la visualización del código

https://github.com/alorher/dataVisualizationPRA2.git

Descripción del tema presentado

Las vacunas protegen a muchas personas de enfermedades peligrosas, pero las vacunas, al igual que las drogas, pueden causar efectos secundarios, un pequeño porcentaje de los cuales pueden ser graves. Para monitorizar estos efectos se creo VAERS (*Vaccine Adverse Event Reporting System*). El *dataset* se compone de tres tablas que se complementan.

Este dataset está elaborado por dos organismos gubernamentales, FDA (Food and Drug Administration) y CDC (Centers for DISEASE Control), hay una mayor confianza en la veracidad de los datos.

La razón de elección de este dataset se justifica en el capítulo siguiente.

Descripción del documento

Este documento pretende servir de justificación de la idoneidad de la visualización, así como del tema elegido. Explicando las razones por las que se ha elegido este tema y de qué forma responde a las preguntas planteadas.

Así mismo, también justifica las razones del diseño y señala los medios empleados y las transformaciones realizadas.

Este proyecto incluye los siguientes elementos:

- Documento pdf con toda la información.
- Libro de trabajo empaquetados de Tableau.
- Script en R con el análisis previo.

EXPLICACIÓN DE LA VISUALIZACION

Interés de la visualización y personas interesadas

La relevancia de este conjunto de datos queda justificada por el marco mundial en el que nos englobamos, la pandemia mundial del COVID-19. El escenario en el que actualmente nos encontramos es el de la "vacunación mundial", llevamos meses dentro de este escenario y ya hemos podido recopilar datos y analizar efectos adversos. Estos datos se recopilan dentro de este dataset. Por todo ello, entre todos los dataset barajados, se ha considerado que este es el más interesante, por los siguientes puntos:

- Es muy actual. Los datos son muy recientes y el escenario es el día a día.
- Gran interés. Las conclusiones que se pueden extraer de este dataset son de gran interés para toda la sociedad, no solo para un pequeño grupo de expertos. Es decir, se podría decir que el colectivo al que afecta es la población mundial
- Gran utilidad. Los datos recopilados, tras su análisis, sirven para establecer futuras estrategias de actuación y entender el paradigma actual de la vacunación, ¿Debemos vacunarnos?
- Perspectiva de género. Los datos podrán tenerse en cuenta tanto para hombres como para mujeres.

Con ello, también resulta interesante estudiar este dataset, ya que, como estamos viendo, las vacunas para el COVID-19 están teniendo unos efectos más adversos de lo esperado. Además, al tratarse de dos vacunas, es difícil discernir que vacuna causo que efecto y si el efecto está relacionado con alguna de las vacunas, esto hace que aun sea más interesante estudiar esta problemática.

Entre el público interesado en la información proporcionada por este *dataset*, tendríamos a los siguientes:

- **Gobiernos de países**. A cualquier gobierno le interesa que ratio de defunciones o efectos adversos produce cada vacuna, con el objetivo de adquirir la de menor ratio. Así como, en el caso de EEUU, le interesa saber las estadísticas por estado, sexo y edad para elaborar informes detallados al respecto y obtener conclusiones.

- **Público en general**. Al ser un tema de actualidad y que nos influye a todos, cualquier persona querrá saber que ocurre con cada vacuna; que efectos y mortalidad tienen.
- **Medios de información**. Por la misma razón que el punto anterior, se trata de un tema de actualidad y que afecta a todo el mundo.
- **Personal sanitario**. Conociendo los efectos adversos de cada vacuna en función del fabricante, sexo y grupo de edad, pueden informar mejor a los pacientes y aplicar la mejor vacuna en función de los criterios anteriores.

Preguntas a responder

Las preguntas a responder se dividen en tres bloques:

Bloque A. Preguntas relativas al número de vacunas

- ***¿A que grupos de edad se ha aplicado el mayor número de vacunas en función del mes y el sexo?***
- ***¿En qué estados se ha aplicado un mayor número, también?***
- ***¿Qué estadísticas hay para cada estado?***

Bloque B. Preguntas relativas a las defunciones

- ***¿En qué grupo de edad, discriminando por sexo se han ocasionados más muertes en el mes de febrero?***
- ***¿Y en que Estados?***

Bloque C. Preguntas relativas al fabricante

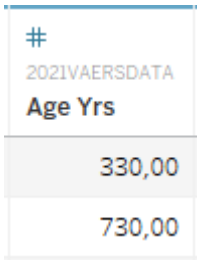
- ***¿Cuáles son los fabricantes más relevantes de vacunas?***
- ***¿Cuáles son los fabricantes con un mayor porcentaje de mortalidad de la vacuna?***

DESCRIPCIÓN TÉCNICA DEL PROYECTO

Transformaciones de datos

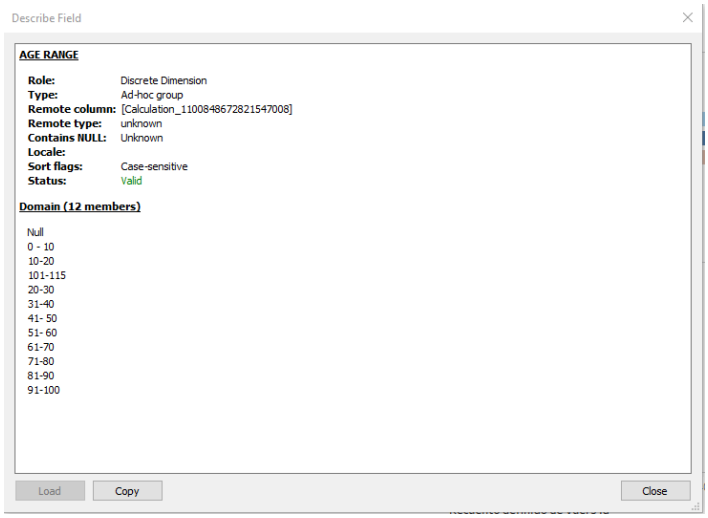
Previamente a la visualización de los datos ha sido necesario realizar un proceso ETL. La extracción de los datos se ha realizado cargándolos directamente de la web *Kaggle* y la carga, es un proceso sencillo en *Tableau*. Por ello, únicamente se comentarán las transformaciones que han sido necesarias realizar, previa a la carga de datos:

- Cambios en tipo de dato.
    - El atributo *Age Yrs* era tipo "string", para trabajar con él, al representar la edad, fue necesario cambiarlo a "number (decimal)". Al realizar esta transformación, la edad cambiaba, al eliminar la coma del decimal, por lo que fue necesario crear otra columna, que sería el valor de esta dividido por 10.

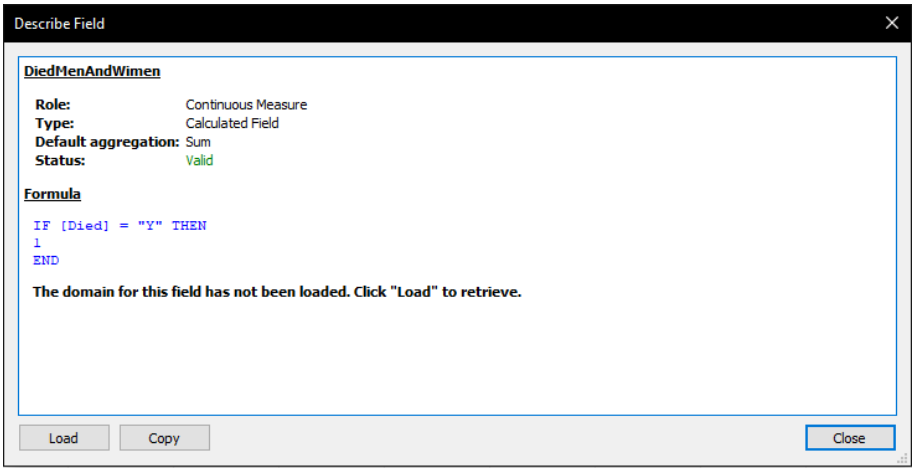


    - La fecha se encontraba en formato MM/DD/YYYY y para trabajar con ella, es necesario tenerla en formato DD/MM/YYYY. Por ello ha sido necesario realizar esta transformación. Con ello, también se han eliminado aquellos registros que tenían esta fecha como "null", así como, aquellos que tenían el atributo VAERS_ID como "null".

- Se ha discretizado la variable AGE, para crear rangos de edad que permitan agrupar las estadísticas por edades de diez en diez:

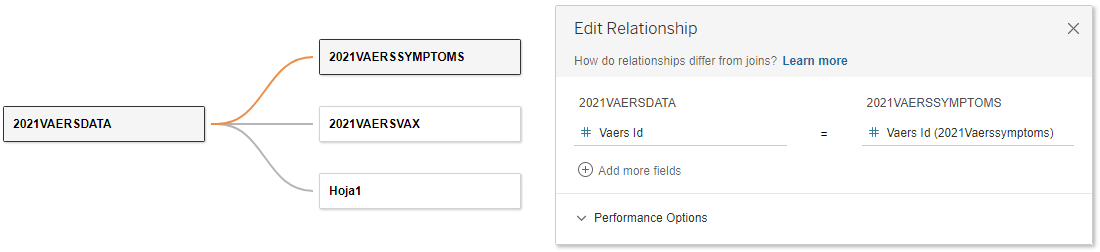


- Se ha creado la columna *DiedMenAndWomen* para indicar mediante un booleano cuando el paciente ha fallecido o no.

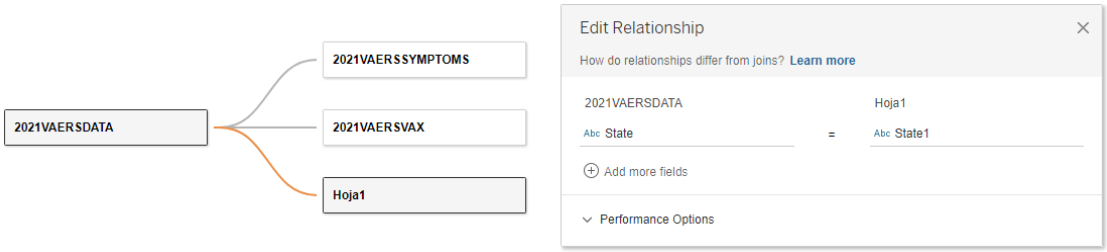


- Se ha creado la columna MaleAndDied, para indicar mediante un booleano si el paciente es hombre y murió.

- **Relación de tablas.** El dataset se compone de tres tablas. Las tres están relacionadas entre si mediante la clave primaria Vaers Id.



Con ello, se ha creado una cuarta tabla con la descripción del código de los estados de EEUU, para facilitar la visualización, cuya clave primaria es State.



Medios técnicos empleados

Para la implementación se han empleado dos herramientas, *Tableau* para la visualización y *R Studio* para el análisis previo de los datos.

R Studio

Se ha empleado R Studio para el proceso de transformación previa a la visualización, con ello, se han realizado las siguientes tareas:

- Comprobación de la estructura de datos.

```
'data.frame':   34121 obs. of  35 variables:
 $ VAERS_ID    : int  916600 916601 916602 916603 916604 9
 $ RECVDATE    : Factor w/ 78 levels "01/01/2021","01/02/2
 $ STATE       : Factor w/ 61 levels "","AK","AL","AR",..:
 $ AGE_YRS     : num  33 73 23 58 47 44 50 33 71 18 ...
 $ CAGE_YR     : int  33 73 23 58 47 44 50 33 71 18 ...
 $ CAGE_MO     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ SEX         : Factor w/ 3 levels "F","M","U": 1 1 1 1 1
 $ RPT_DATE    : Factor w/ 25 levels "","01/02/2021",..: 1
 $ SYMPTOM_TEXT: Factor w/ 33194 levels "'AT APPROX 1007AM
 19137 20274 14057 ...
 $ DIED        : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ DATEDIED    : Factor w/ 103 levels "","01/01/2021",..:
 $ L_THREAT    : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ ER_VISIT    : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ HOSPITAL    : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ HOSPDAYS    : int  NA NA NA NA NA NA NA NA NA NA ...
 $ X_STAY      : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ DISABLE     : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ RECOVD      : Factor w/ 4 levels "","N","U","Y": 4 4 3
 $ VAX_DATE    : Factor w/ 297 levels "","01/01/1921",..:
 $ ONSET_DATE  : Factor w/ 221 levels "","01/01/2011",..:
 $ NUMDAYS     : int  2 0 0 0 7 0 1 2 8 1 ...
 $ LAB_DATA    : Factor w/ 8913 levels "","-","--","---
 4152 1 ...
 $ V_ADMINBY   : Factor w/ 10 levels "","MIL","OTH",..: 6
 $ V_FUNDBY    : Factor w/ 5 levels "","OTH","PUB",..: 1 1
 $ OTHER_MEDS  : Factor w/ 14978 levels "","-","- Effexor
 $ CUR_ILL     : Factor w/ 4213 levels "","-","--","- covi
 $ HISTORY     : Factor w/ 11527 levels "","-","- colon ca
535 3990 1 ...
 $ PRIOR_VAX   : Factor w/ 1319 levels "","\"flushing,\" 5
 $ SPLTTYPE    : Factor w/ 6460 levels "","?","000027",..:
 $ FORM_VERS   : int  2 2 2 2 2 2 2 2 2 2 ...
 $ TODAYS_DATE : Factor w/ 113 levels "","01/01/2020",..:
 $ BIRTH_DEFECT: Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1
 $ OFC_VISIT   : Factor w/ 2 levels "","Y": 2 2 1 1 1 1 1
 $ ER_ED_VISIT : Factor w/ 2 levels "","Y": 1 1 2 1 1 1 1
 $ ALLERGIES   : Factor w/ 7750 levels "","-","----",..: 4
```

- Verificación de la idoneidad del nombre de los atributos, se ha comprobado si era necesario cambiar el nombre de algún atributo para evitar problemas. No ha sido necesario cambiar ningún nombre
- Extracción de estadísticas básicas, para tener una orientación sobre la futura visualización.
- Comprobación de valores vacíos

VAERS_ID	RECVDATE	STATE	AGE_YRS	CAGE_YR	CAGE_MO	SEX	RPT_DATE	SYMPTOM_TEXT	DIED	DATEDIED	L_THREAT	ER_VISIT	HOSPITAL	HOSPDAYS
0	0	5571	NA	NA	NA	0	34058	0	32168	32327	32865	34110	29742	NA
X_STAY	DISABLE	RECOVD	VAX_DATE	ONSET_DATE	NUMDAYS	LAB_DATA	V_ADMINBY	V_FUNDBY	OTHER_MEDS	CUR_ILL	HISTORY	PRIOR_VAX	SPLTTYPE	FORM_VERS
34069	33254	2892	1538	1901	NA	NA	53	34057	NA	NA	NA	32690	NA	NA
TODAYS_DATE	BIRTH_DEFECT	OFC_VISIT	ER_ED_VISIT	ALLERGIES										
282	34070	28735	28614	NA										

- Cambio de los valores vacíos de STATE por “Desconocido”.

Con ello, se han cargado la librería *dplyr* y *ggplot2*. Aunque esta última, meramente para crear visualizaciones previas para tener una idea de los datos. Estos gráficos no se incluyen, ya que no son relevantes una vez desarrollados los dashboards.

Tableau

Se han creado 3 dashboards distintos para responder a distintas preguntas. Esta herramienta dada su alta funcionalidad también permite la manipulación de distintos elementos de la visualización.

Se ha elegido este software por varios motivos:

- Cuenta con numerosas herramientas de visualización: mapas, gráficos, *treemaps*, *packed bubbles*…
- La interfaz es muy amigable e intuitiva
- Permite correlacionar varias tablas muy fácilmente
- Permite transformaciones de datos una vez cargados.
- Es adaptable a distintas situaciones y propósitos.
- Es compatible con un gran número de fuentes de datos.
- No es necesario programar, únicamente al crear alguna función, con un código muy sencillo, como un “if”.
- Es rápido, la carga de datos y procesamiento lleva poco tiempo.

VISUALIZACIÓN DE LOS DATOS

Justificación de idoneidad

Se han creado 3 dashboards para responder a las cuestiones anterioes, justificando asi, que la visualización es correcta, al proporcionar toda la información que necesitamos, respondiendo a las preguntas planteadas en el Capítulo 2.

Dashboard del número de vacunas

Este *dashboard* representa el número de vacunas aplicadas y se muestra de las siguientes formas:

- En un mapa de EEUU con el número de vacunas aplicadas en cada estado,
- un gráfico de doble eje que combina el rango de edad con el sexo, y
- en un gráfico de burbujas que presenta el número de vacunas aplicadas por mes

Junto a estos gráficos, los datos pueden ser manipulados filtrando por sexo y rango temporal. Así mismo, el grafico de burbujas y el gráfico con los rangos de edad también sirve como filtros.

Dashboard del número de muertes

Este *dashboard* representa el número de muertes producidas tras la aplicación de la vacuna y se muestra de las siguientes formas:

- En un mapa de EEUU con el número de muertes en cada estado,
- Un gráfico de barras indicado el número de muertes para cada grupo de edad y diferenciando por sexo,
- Un gráfico circular indicando el número de muertes y el porcentaje respecto al total para cada sexo

Junto a estos gráficos, los datos pueden ser manipulados filtrando por sexo y rango temporal. Así mismo, el grafico de burbujas y el gráfico con los rangos de edad también sirve como filtros.

Dashboard de datos de los fabricantes

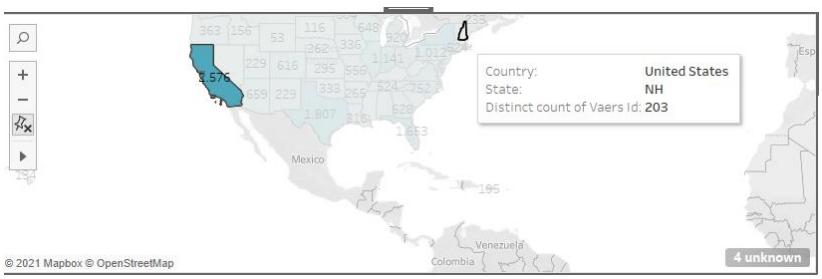
Este *dashboard* representa el número de vacunas aplicadas y el número de muertes producidas para cada fabricante de la vacuna. Los datos se muestran de la siguiente manera:

- Un gráfico de burbujas indicando el número de vacunas y el porcentaje respecto al total.
- Para clarificar el grafico de burbujas se ha incluido una lista con el número de vacunas por fabricante.
- Un "treemap" indicando la mortalidad de la vacuna.
- Un gráfico de burbujas indicando el número de vacunas aplicadas por fabricante.

Justificación de diseño

El diseño queda justificado por las siguientes razones:

- Se han creado 3 dashboards en función de los objetivos de las agrupaciones de las preguntas: número de vacunas, defunciones y fabricantes.
- Los *dahsboards* permiten la iteración con los datos de distintas formas:
    - Mediante los filtros que permiten manipular la visualización de los datos en función de nuestras preferencias. Podemos filtrar por sexo y por marco temporal.
    - Haciendo "click" en cualquiera de los elementos de la visualización, podemos obtener unos datos concretos. Por ejemplo, si queremos ver las estadísticas de cada estado, haciendo "click", en el estado en cuestión, podremos ver sus estadísticas, manipulando el resto de los elementos de la visualización.



- El tamaño de los elementos de la visualización permite ver los datos sin ningún tipo de esfuerzo y sin que sea excesivamente grande.

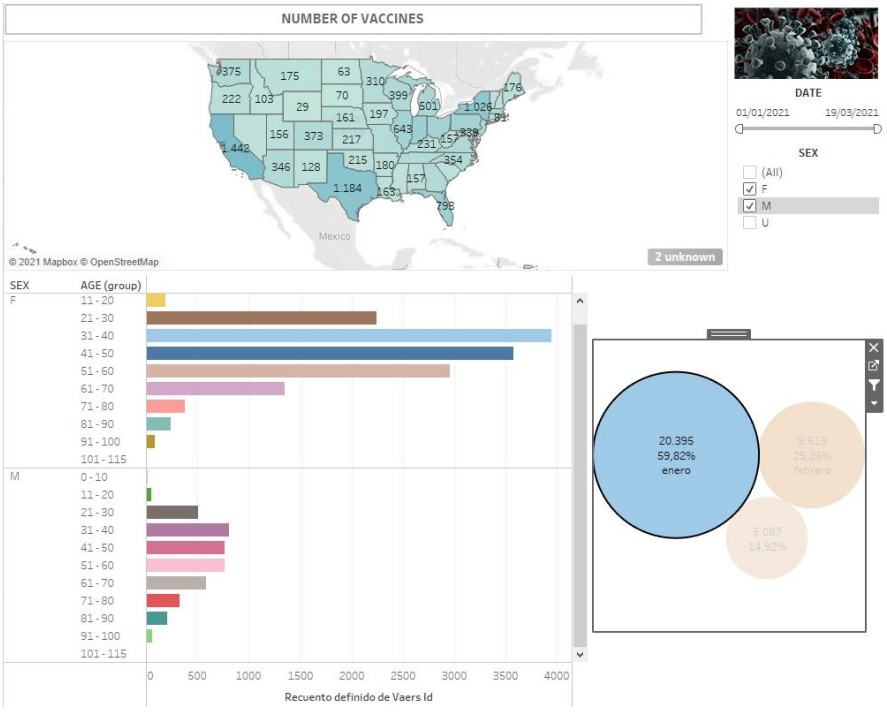
- Los colores son tenues y se encuentran todos en la misma tonalidad, de manera que no cansan a la vista y no hay despuntes de tonalidad que puedan confundir o sean estéticamente desagradables.
- Los títulos de cada dashboards son cortos, concretos y definen la visualización sin ambigüedades.

Respuestas a las preguntas planteadas

Los dashboard anteriores se han creado para dar respuesta a los siguientes tres grupos de preguntas:

¿A que grupos de edad se ha aplicado el mayor número de vacunas en función del mes y el sexo? ¿En qué estados se ha aplicado un mayor número, también? ¿Qué estadísticas hay para cada estado?

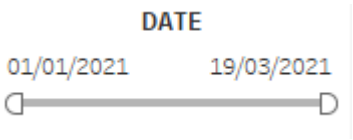
Para responder a esta pregunta deberemos elegir el Dasboard "Vaccines" y en el gráfico de burbujas seleccionar un mes, por ejemplo, el mes de enero:



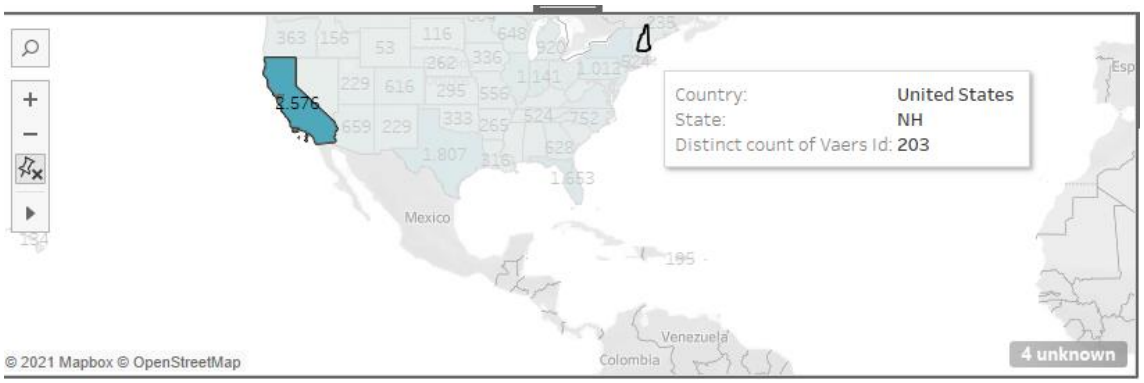
Para el mes de enero, vemos que se han aplicado 20.9395 vacunas, que supone un 59.82% del primer trimestre. Así mismo, vemos que el mayor grupo al que se ha aplicado esta vacuna ha sido al grupo de 31-40 años, seguido de 41-50 y después del grupo 51-60. Esto

tiene sentido, ya que, en este mes, comenzó la vacunación y en primer lugar se vacuno al personal sanitario, que se encuentra en estos rangos de edad. Finalmente, vemos, que el estado en el que se vacuno más, fue California.

Lo anterior, también podría haberse obtenido filtrando con el filtro DATE:

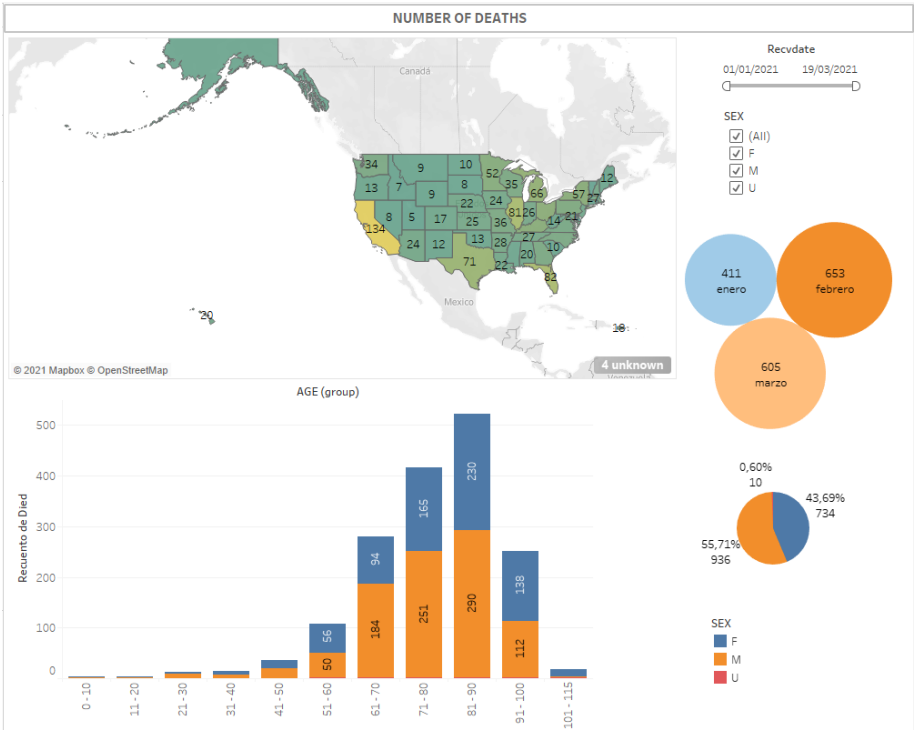


Para obtener las estadísticas de cada estado, únicamente hay que pinchar en ele stado en cuestión y obtendremos estas estadísticas, por ejemplo, California:



¿En qué grupo de edad, discriminando por sexo se han ocasionados más muertes en el mes de febrero? ¿Y en que Estados?

Para responder a lo anterior empleamos el Dashboard “Deaths”. Seleccionamos el gráfico de burbujas el mes de febrero y obtenemos lo siguientes:



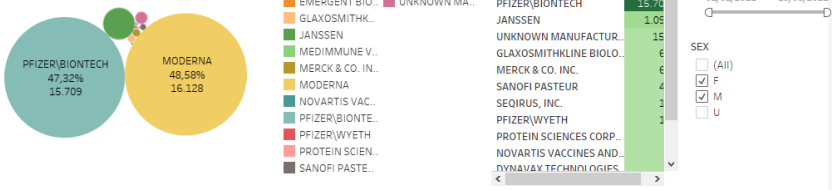
El mayor número de muertes se ha producido para los rangos de edad de 81-90 años, seguido de 71-80 años. Siendo el número de varones ligeramente superior al de mujeres. Además, vemos, que el estado con más defunciones ha sido California.

¿Cuáles son los fabricantes más relevantes de vacunas? ¿Cuáles son los fabricantes con un mayor porcentaje de mortalidad de la vacuna?

Para responder a lo anterior empleamos el Dashboard “Manufacturer”.

Los fabricantes más relevantes son PFIZER y MODERNA. Entre los dos, tienen más del 90% del mercado. La mortalidad de PZIER es del 4.45% (725 muertes) y de MODERNA (895 muertes)5.47%. Llama la atención el 100% de mortalidad de DYNAVAX, pero esto es porque únicamente se aplicó una vacuna, frente a las 16.374 vacunas de MODERNA y 16.283 DE PZIZER.

VACCINES PER MANUFACTER



DEATHS PER MANUFACTURER