# Bootcamp: Análisis y Visualización de Datos

## Código: AVDV2-47
## Nivel intermedio

## Integrantes

Aldemar Bohórquez
José Valdés Domínguez
Jair Díez Henao

## Equipo 5

DATAWI
DECISIONES BASADAS EN DATOS

# Índice

# Cronograma de actividades

Día 1: Comprensión del negocio | 3 horas

Reunión inicial con el equipo para definir objetivos del proyecto.
Identificación y documentación de los requisitos del negocio.
Redacción del plan de proyecto y establecimiento de métricas clave de éxito.

Día 2, 3 y 4: Preparación de los datos | 9 horas

Limpieza de datos (manejo de valores nulos, duplicados y errores).
Investigación de la naturaleza de las variables.
Transformación de datos (conversión de tipos de datos, verificación y contraste con otras bases)

Día 5: Comprensión de los datos | 3 horas

Descarga del conjunto de "Casos positivos de COVID-19 en Colombia".
Exploración inicial de los datos, revisión de la estructura y tipos de datos.
Identificación de posibles problemas de calidad de los datos y plan para resolverlos.

Día 6 y 7: Exploración de datos | 6 horas

Exploración de los datos y creación de variables en caso de ser necesario para un análisis profundo.
Análisis descriptivo univariado de las variables finales.

Día 8 : Evaluación | 3 horas

Evaluación de las métricas pertinentes para un tablero de POWER BI
Revisión de todas las etapas del proyecto para asegurarse de que se han completado correctamente.
Decisión sobre las métricas finales y preparación de los resultados para la implementación.

Día 9 y 10: Despliegue y visualización en Power BI | 6 horas

Preparación de los datos finales para la visualización en Power BI.
Diseño de dashboards en Power BI, creación de gráficos y visualizaciones necesarias.

Preparación de la presentación final, incluyendo la narrativa del análisis.
Ensayo de la presentación y ajustes finales.

Día 11: Presentación final | 1 hora

Revisión final de todos los materiales y dashboards en Power BI.
Presentación final del proyecto a los stakeholders.

# Fase 1: comprensión del negocio

DATAWI es una empresa colombiana dedicada a la prestación de servicios en análisis, visualización y minería de datos, inteligencia de negocio, analítica digital avanzada y desarrollo de software a medida, para las empresas públicas y privadas en todos los sectores empresariales.

En el marco del presente proyecto, DATAWI fue contratada por el gobierno colombiano con el objetivo de desarrollar un tablero público interactivo en la web. Esta herramienta permite explorar la totalidad de los datos recopilados sobre el comportamiento de la pandemia COVID-19 en Colombia, desde su declaración el 13 de marzo de 2020 hasta su finalización el 20 de junio de 2022. Cabe destacar que los datos disponibles se extienden hasta el 6 de enero de 2024.

## Determinación de los objetivos del proyecto de Ciencia de Datos

**Objetivo**: Proporcionar un tablero interactivo en Power BI que presente de manera entendible, detallada y descriptiva, la información recopilada sobre la pandemia del COVID-19 en la República de Colombia, el cual permita a las entidades gubernamentales y privadas realizar análisis retrospectivos del comportamiento del virus en la población.

- *Objetivo específico 1*: Realizar limpieza y puesta a punto de todos los datos entregados por el gobierno colombiano.
- *Objetivo específico 2*: Proveer análisis descriptivo de las variables contenidos en el conjunto de datos.
- *Objetivo específico 3*: Desarrollar y desplegar un tablero interactivo de Power BI para la visualización de los datos.

# Fase 2: preparación de los datos

En la siguiente tabla se encuentra una descripción de los cambios realizados para la limpieza de los datos.

- El conjunto de datos inicial tiene un total de 6.390.971 filas y 23 columnas: Tipos de datos: float64(2), int64(5), object(16)

| | NOMBRE COLUMNA | TIPO |
|---|---|---|
| 0 | fecha reporte web | object |
| 1 | ID de caso | int64 |
| 2 | Fechas de notificación | object |
| 3 | Código DIVIPOLA departamento | int64 |
| 4 | Nombre departamento | object |
| 5 | Código DIVIPOLA municipio | int64 |
| 6 | Nombre municipio | object |
| 7 | Edad | int64 |
| 8 | Unidad de medida de edad | int64 |
| 9 | Sexo | object |
| 10 | Tipo de contagio | object |
| 11 | Ubicación del caso | object |
| 12 | Estado | object |
| 13 | Código ISO del país | float64 |
| 14 | Nombre del país | object |
| 15 | Recuperado | object |
| 16 | Fecha de inicio de síntomas | object |
| 17 | Fecha de muerte | object |
| 18 | Fecha de diagnóstico | object |
| 19 | Fecha de recuperación | object |
| 20 | Tipo de recuperación | object |
| 21 | Pertenencia étnica | float64 |
| 22 | Nombre del grupo étnico | object |

- La siguiente tabla muestra la cantidad de valores nulos por cada una de las columnas:

| Cantidad_Nulos | Porcentajes_Nulos | |
|---|---|---|
| fecha reporte web | 0 | 0 |
| ID de caso | 0 | 0 |
| Fecha de notificación | 0 | 0 |
| Código DIVIPOLA departamento | 0 | 0 |
| Nombre departamento | 0 | 0 |
| Código DIVIPOLA municipio | 0 | 0 |
| Nombre municipio | 0 | 0 |
| Edad | 0 | 0 |
| Unidad de medida de edad | 0 | 0 |
| Sexo | 0 | 0 |
| Tipo de contagio | 0 | 0 |
| Ubicación del caso | 41,268 | 0.65 |
| Estado | 41,268 | 0.65 |
| Código ISO del país | 6,387,261 | 99.94 |
| Nombre del país | 6,387,253 | 99.94 |
| Recuperado | 36,327 | 0.57 |
| Fecha de inicio de síntomas | 505,384 | 7.91 |
| Fecha de muerte | 6,206,578 | 97.11 |
| Fecha de diagnóstico | 2,755 | 0.04 |
| Fecha de recuperación | 182,324 | 2.85 |
| Tipo de recuperación | 182,398 | 2.85 |
| Pertenencia étnica | 2,156 | 0.03 |
| Nombre del grupo étnico | 6,307,107 | 98.69 |

Las columnas con más cantidad de nulos son “Código ISO del país” y “Nombre del país” (99.94%), seguidas por “Nombre del grupo étnico” (98.69%) y “Fecha de muerte” (97.11%).

- Se observa que la información proporcionada por las siguientes columnas es redundante y/o no aporta información relevante para nuestro proyecto:

- Fecha de reporte web
- ID del caso
- Tipo de recuperación
- Nombre del grupo étnico

En las siguientes columnas se evidencia errores en los códigos:

- “Código DIVIPOLA departamento” (47001, 8001, 13001), generando error también en el nombre de Nombre departamento.
- “Código DIVIPOLA municipio”.

Además, se deben unificar/estandarizar los valores o categorías en las siguientes columnas:

- Nombre departamento
- Nombre municipio
- Sexo
- Ubicación del caso
- Estado
- Nombre del país
- Recuperado
- Tipo de recuperación

Adicionalmente, se detectaron errores en las columnas “Unidad de medida de edad”, pues las categorías 5 y 4 no forman parte del diccionario de datos, “código ISO del país”, dado que estos códigos no pueden ser de 4 cifras, y “Pertenencia étnica”, donde no deben existir valores nulos de acuerdo al diccionario de datos.

## Limpieza de los datos

1. Se procede a eliminar las categorías 4 y 5 de la columna “Unidad de medida de edad”, puesto que solo hay un dato de cada una y en el conjunto de datos original no existe explicación para esas dos categorías.
2. Corrección de errores en las columnas “Código DIVIPOLA departamento” y “Código DIVIPOLA municipio”.

Inicialmente se realizará un reemplazo de los códigos que no corresponden a cada columna, posteriormente se cargará un dataset que contenga los códigos DIVIPOLA de departamentos y municipios para hacer una unión con el dataframe principal (df_casos_covid_colombia).

Inicialmente se realizará un reemplazo de los códigos que no corresponden a cada columna, posteriormente se cargará un dataset que contenga los códigos DIVIPOLA de departamentos y municipios para hacer una unión con el dataframe principal (df_casos_covid_colombia).

3. Corrección de errores en la columna “Código ISO del país”: al igual que en las columnas anteriores, inicialmente se realizará un reemplazo de los códigos ISO que no corresponden a un país, posteriormente se cargará un dataset que contenga dichos códigos para hacer una unión con el dataframe principal (df_casos_covid_colombia).

4. Imputación de los valores nulos en la columna “NOMBRE_PAIS”: como esta columna se refiere al país de donde proviene la persona infectada entonces los valores nulos indican que la persona se contagió en Colombia.

5. Corrección/Imputación de los valores nulos en la columna “Pertenencia étnica”: debido a que no deben existir valores nulos en la columna, se imputarán estos valores por el número 7 aludiendo a una nueva categoría denominada 'No_aplica'.

6. Corrección/Imputación de los valores nulos en la columna “Unidad de medida de edad”: esto para quitar los acentos (1:'Anios', 2:'Meses', 3:'Dias').

7. Cambiar el tipo de dato *object* de las columnas que tienen fechas: “Fecha de notificación”, “Fecha de inicio de síntomas”, “Fecha de muerte”, “Fecha de diagnóstico” y “Fecha de recuperación”.

8. Corrección/Imputación de los valores nulos en la columna “ubicación del caso”, al analizar esta columna se detecta: ['Casa' 'Fallecido' 'Desconocido' 'casa' 'CASA' 'Hospital' 'Hospital UCI'].

9. Corrección/Imputación de los valores nulos en la columna “Estado”, al analizar esta columna se detecta: ['Leve' 'Fallecido' 'Desconocido' 'leve' 'LEVE' 'Moderado' 'Grave'].

10. Corrección/Imputación de los valores nulos en la columna “Recuperado”, al analizar esta columna se detecta: ['Recuperado' 'Fallecido' 'Fallecido NO COVID' 'fallecido' 'Activo'].

11. Se procede a eliminar columna que contienen información redundante: [“fecha reporte web”, “ID de caso”, “Código DIVIPOLA departamento”, “Nombre departamento”, “Código DIVIPOLA municipio”, “Nombre municipio”, “Tipo de recuperación”, “Código ISO del país”, “Nombre del país”, “Código Departamento”, “Código Municipio”, “Nombre del grupo étnico”, “CODIGO_ISO_PAIS”].

12. Unificar el nombre de las columnas a mayúsculas y sin acentos: [“FECHA_DE_NOTIFICACION”, “EDAD”, “UNIDAD_DE_MEDIDA_DE_EDAD”, “SEXO”, “TIPO_DE_CONTAGIO”, “UBICACION_DEL_CASO”, “ESTADO”, “RECUPERADO”, “FECHA_DE_INICIO_DE_SINTOMAS”, “FECHA_DE_MUERTE”, “FECHA_DE_DIAGNOSTICO”, “FECHA_DE_RECUPERACION”, “PERTENENCIA_ETNICA”, “NOMBRE_DEPARTAMENTO”, “NOMBRE_MUNICIPIO”, “NOMBRE_PAIS”].

13. Estandarizar/Unificar las categorías de las columnas:
    Columna unificada: SEXO
    Columna unificada: UBICACION_DEL_CASO
    Columna unificada: ESTADO
    Columna unificada: RECUPERADO
    Columna unificada: NOMBRE_DEPARTAMENTO
    Columna unificada: NOMBRE_MUNICIPIO
    Columna unificada: NOMBRE_PAIS

14. Información actual del DataFrame: el Dataframe tiene 6.390.969 filas y 16 columnas.

| | Cantidad_Nulos | Porcentajes_Nulos |
|---|---|---|
| FECHA_DE_NOTIFICACION | 0 | 0 |
| EDAD | 0 | 0 |
| UNIDAD_DE_MEDIDA_DE_EDAD | 0 | 0 |
| SEXO | 0 | 0 |
| TIPO_DE_CONTAGIO | 0 | 0 |
| UBICACION_DEL_CASO | 0 | 0 |
| ESTADO | 0 | 0 |
| RECUPERADO | 0 | 0 |
| FECHA_DE_INICIO_DE_SINTOMAS | 505,384 | 7.91 |
| FECHA_DE_MUERTE | 6,206,576 | 97.11 |
| FECHA_DE_DIAGNOSTICO | 2,755 | 0.04 |
| FECHA_DE_RECUPERACION | 182,324 | 2.85 |
| PERTENENCIA_ETNICA | 0 | 0 |
| NOMBRE_DEPARTAMENTO | 0 | 0 |
| NOMBRE_MUNICIPIO | 0 | 0 |
| NOMBRE_PAIS | 0 | 0 |

Se observa la presencia de valores nulos en las columnas con fechas. Para determinar el tratamiento adecuado de estas variables, se consultó el diccionario de datos y se analizaron las relaciones entre ellas. Por ejemplo, si existe la variable “FECHA_DE_MUERTE” y la variable “FECHA_DE_RECUPERACION” presenta valores nulos, se debe revisar esta relación para determinar el manejo adecuado de los campos nulos en ambas variables.

15. Corrección/Imputación de los valores nulos en las columnas “FECHA_DE_INICIO_DE_SINTOMAS” y “FECHA_DE_DIAGNOSTICO”.
Un total de 274,473 casos tienen la misma fecha para ambas columnas, lo cual nos muestra que la “fecha de inicio de sintomas” y la “fecha diagnostico”, la cual es la fecha de confirmación por parte del laboratorio, coinciden en algunos casos.

Además, se observa que del total de valores nulos (505,384) de la columna "FECHA_DE_INICIO_DE_SINTOMAS", 505.282 registros presentan valores en la variable "FECHA_DE_DIAGNOSTICO". Dado que esta última representa la fecha de confirmación por laboratorio, se asumirá que la prueba se realizó por sospecha el mismo día. Por lo tanto, se procederá a llenar los valores nulos de "FECHA_DE_INICIO_DE_SINTOMAS" con las fechas presentes en la variable "FECHA_DE_DIAGNOSTICO".
Para el resto de registros (102), que son nulos para ambas columnas, se toma la decisión de eliminarlas dado que representa apenas el 0.001% de los datos.

16. Ahora verificaremos de nuevo los datos nulos del dataframe completo con los cambios realizados.

| | Cantidad_Nulos | Porcentajes_Nulos |
|---|---|---|
| FECHA_DE_NOTIFICACION | 0 | 0 |
| EDAD | 0 | 0 |
| UNIDAD_DE_MEDIDA_DE_EDAD | 0 | 0 |
| SEXO | 0 | 0 |
| TIPO_DE_CONTAGIO | 0 | 0 |
| UBICACION_DEL_CASO | 0 | 0 |
| ESTADO | 0 | 0 |
| RECUPERADO | 0 | 0 |
| FECHA_DE_INICIO_DE_SINTOMAS | 0 | 0 |
| FECHA_DE_MUERTE | 6,206,474 | 97.11 |
| FECHA_DE_DIAGNOSTICO | 2,653 | 0.04 |
| FECHA_DE_RECUPERACION | 182,324 | 2.85 |
| PERTENENCIA_ETNICA | 0 | 0 |
| NOMBRE_DEPARTAMENTO | 0 | 0 |
| NOMBRE_MUNICIPIO | 0 | 0 |
| NOMBRE_PAIS | 0 | 0 |

Ahora, para los registros nulos que tiene la columna "FECHA_DE_DIAGNOSTICO" se tomarán las fechas registradas como inicio de síntomas que las proporciona la columna "FECHA_DE_INICIO_DE_SINTOMAS".

| | Cantidad_Nulos | Porcentajes_Nulos |
|---|---|---|
| FECHA_DE_NOTIFICACION | 0 | 0 |
| EDAD | 0 | 0 |
| UNIDAD_DE_MEDIDA_DE_EDAD | 0 | 0 |
| SEXO | 0 | 0 |
| TIPO_DE_CONTAGIO | 0 | 0 |
| UBICACION_DEL_CASO | 0 | 0 |
| ESTADO | 0 | 0 |
| RECUPERADO | 0 | 0 |
| FECHA_DE_INICIO_DE_SINTOMAS | 0 | 0 |
| FECHA_DE_MUERTE | 6,206,474 | 97.11 |
| FECHA_DE_DIAGNOSTICO | 0 | 0 |
| FECHA_DE_RECUPERACION | 182,324 | 2.85 |
| PERTENENCIA_ETNICA | 0 | 0 |
| NOMBRE_DEPARTAMENTO | 0 | 0 |
| NOMBRE_MUNICIPIO | 0 | 0 |
| NOMBRE_PAIS | 0 | 0 |

17. Análisis y corrección/imputación de los registros en las columnas “FECHA_DE_MUERTE”, “FECHA_DE_RECUPERACION”, “RECUPERADO”. Dada la conexión existente entre estas dos variables miraremos para cuántos registros no se tiene información de ninguna de las dos, para un total de 2,946. Debido a que es posible que para estos casos no se tenga un seguimiento, pues solo se sabe que tuvieron síntomas y un diagnóstico pero no lo qué sucedió con el caso, y además representan un 0.046% de los datos actuales se procede a eliminar dichos registros.

Volvemos a verificar los datos nulos del dataframe completo con los cambios realizados:

| | Cantidad_Nulos | Porcentajes_Nulos |
|---|---|---|
| FECHA_DE_NOTIFICACION | 0 | 0 |
| EDAD | 0 | 0 |
| UNIDAD_DE_MEDIDA_DE_EDAD | 0 | 0 |
| SEXO | 0 | 0 |
| TIPO_DE_CONTAGIO | 0 | 0 |
| UBICACION_DEL_CASO | 0 | 0 |
| ESTADO | 0 | 0 |
| RECUPERADO | 0 | 0 |
| FECHA_DE_INICIO_DE_SINTOMAS | 0 | 0 |
| FECHA_DE_MUERTE | 6,203,528 | 97.11 |
| FECHA_DE_DIAGNOSTICO | 0 | 0 |
| FECHA_DE_RECUPERACION | 179,378 | 2.81 |
| PERTENENCIA_ETNICA | 0 | 0 |
| NOMBRE_DEPARTAMENTO | 0 | 0 |
| NOMBRE_MUNICIPIO | 0 | 0 |
| NOMBRE_PAIS | 0 | 0 |

18. Ahora miraremos la conexión existente entre “FECHA_DE_MUERTE” y “FECHA_DE_RECUPERACION”.

    Observamos que del total de registros, que son (6,387,921), existen 6,203,528 registros donde no se tiene “FECHA_DE_MUERTE” pero sí “FECHA_DE_RECUPERACION”, 179,378 registros donde no se tiene “FECHA_DE_RECUPERACION” pero sí “FECHA_DE_MUERTE” y 5,015 registros donde ambas columnas existen (***¡¡¡un caso particular!!!***) que se verifica nuevamente pero mostrando la diferencia de días entre estas dos columnas, descartando así que “FECHA_DE_MUERTE” sea menor que “FECHA_DE_RECUPERACION”.

    Con lo anterior corrobora que “FECHA_DE_MUERTE” es igual (0 días) o mayor que “FECHA_DE_RECUPERACION”. Seguidamente, se debe verificar entonces

que estos 5,015 registros aparezcan como Fallecido en la columna “RECUPERADO”.

Recuperado = 4941
Fallecido_No_Covid = 74

19. Se observa que no es el caso para 4,941 registros, por lo tanto, se procede a imputar dichos valores.

20. Identificación del data frame final: El Dataframe final tiene 6387921 filas y 16 columnas.

| | Cantidad_Nulos | Porcentajes_Nulos |
|---|---|---|
| FECHA_DE_NOTIFICACION | 0 | 0 |
| EDAD | 0 | 0 |
| UNIDAD_DE_MEDIDA_DE_EDAD | 0 | 0 |
| SEXO | 0 | 0 |
| TIPO_DE_CONTAGIO | 0 | 0 |
| UBICACION_DEL_CASO | 0 | 0 |
| ESTADO | 0 | 0 |
| RECUPERADO | 0 | 0 |
| FECHA_DE_INICIO_DE_SINTOMAS | 0 | 0 |
| FECHA_DE_MUERTE | 6,203,528 | 97.11 |
| FECHA_DE_DIAGNOSTICO | 0 | 0 |
| FECHA_DE_RECUPERACION | 179,378 | 2.81 |
| PERTENENCIA_ETNICA | 0 | 0 |
| NOMBRE_DEPARTAMENTO | 0 | 0 |
| NOMBRE_MUNICIPIO | 0 | 0 |
| NOMBRE_PAIS | 0 | 0 |

21. Se procede a exportar el conjunto de datos final para usarlo en Power BI y continuar con la construcción y diseño del tablero solicitado.

# Fase 3: Comprensión de los datos

## Información sobre el conjunto de datos:

**Nombre**: Casos positivos de COVID-19 en Colombia
**Fecha de creación**: 27 de marzo de 2020
**Propietario de conjunto de datos**: Instituto Nacional de Salud
**Idioma**: Español
**Cobertura geográfica**: Nacional (República de Colombia)
**Frecuencia de actualización**: diaria
**Fecha emisión** (aaaa-mm-dd): 2020-03-27

¿Qué hay en este conjunto de datos?
Filas: 6.387.921 | Columnas: 16

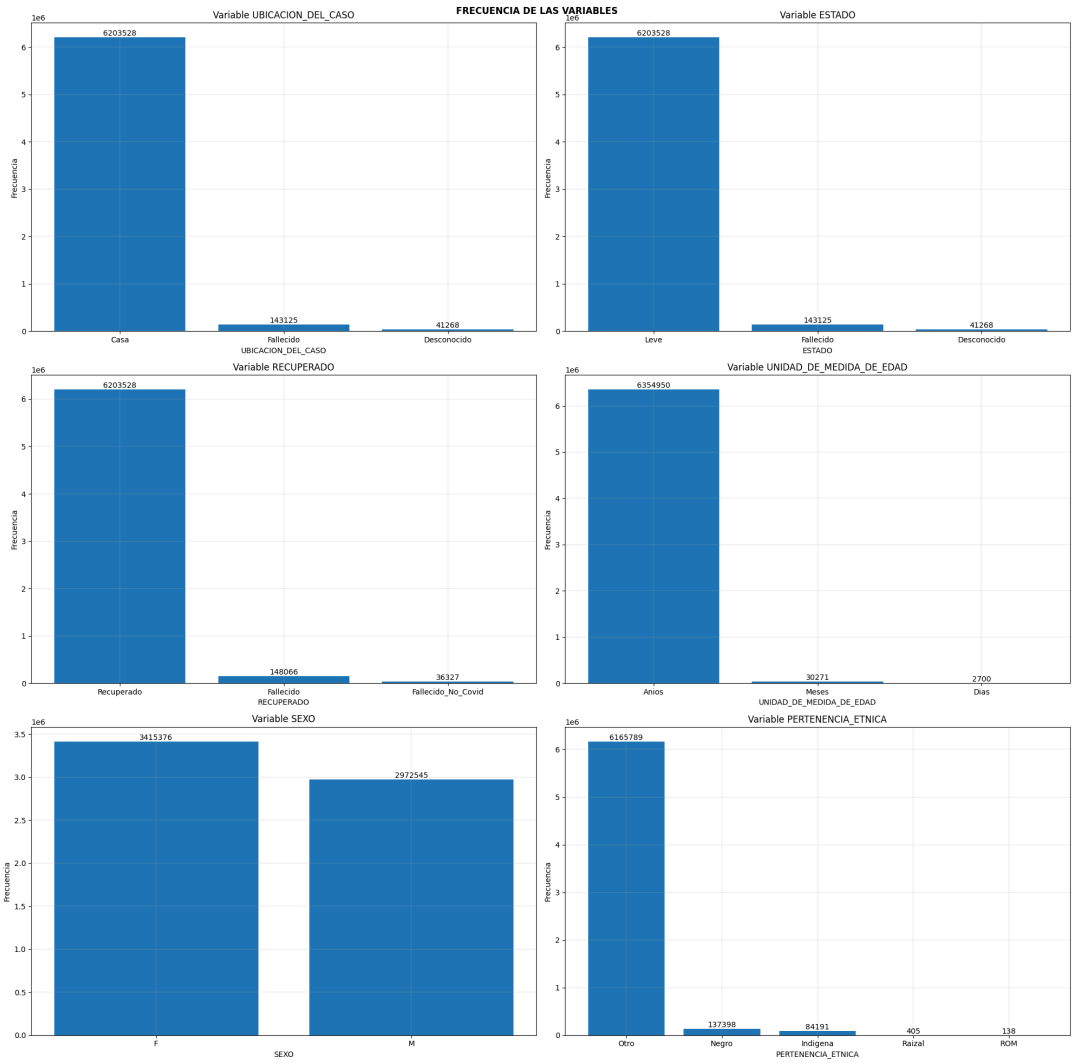
## Diccionario de datos: este es post-limpieza de los datos.

| Nombre | Tipo | Descripción |
|---|---|---|
| EDAD | Entero | Corresponde a la edad de la persona al momento de notificarle positivo para COVID-19. |
| UNIDAD_DE_MEDIDA_DE_EDAD | Texto | Esta corresponde a la unidad de medida para las edades: Anios, Meses y Dias. |
| SEXO | Texto | Corresponde al género de la persona entre M (Masculino) y F (Femenino). |
| TIPO_DE_CONTAGIO | Texto | Es el tipo de contagio que determinan los profesionales de la salud al momento de la notificación. Son 4 tipos: Relacionado, Importado, En estudio o Comunitario. |
| UBICACION_DEL_CASO | Texto | Corresponde a la ubicación del caso indicado por los profesionales de salud y tiene 3 categorías: casa, fallecido y desconocido. |
| ESTADO | Texto | Corresponde al estado del caso y tiene 3 categorías: leve, fallecido y desconocido. |
| FECHA_DE_NOTIFICACION | Fecha | Corresponde a la fecha de notificación entregada al SIVIGILA (Sistema de vigilancia en Salud Pública) que el profesional de la salud le indica. |
| NOMBRE_PAIS | Texto | Corresponde al nombre del país en español. Este es el país de procedencia del posible contagio. |
| RECUPERADO | Texto | Recuperado, Fallecido, N/A, (Vacío). N/A se refiere a los fallecidos no COVID. Pueden haber casos recuperados con ubicación Hospital u Hospital UCI, ya que permanecen en hospitalización por causas diferentes. Los casos con |

| | | |
|---|---|---|
| | | información en blanco en esta columna corresponden a los casos activos. |
| FECHA_DE_INICIO_DE_SINTOMAS | Fecha | Es la fecha en que la persona indica que iniciaron los síntomas de COVID-19. |
| FECHA_DE_MUERTE | Fecha | Fecha en la que se reporta la muerte de la persona por COVID-19. |
| FECHA_DE_DIAGNOSTICO | Fecha | Corresponde a la fecha de confirmación por parte del laboratorio que procesó la muestra. |
| FECHA_DE_RECUPERACION | Fecha | Fecha en la que se da de alta a la persona. |
| PERTENENCIA_ETNICA | Texto | Corresponde a la pertenencia étnica que el profesional de la salud le pregunta a la persona. Estos son los tipos: Indigena, ROM, Raizal, Palenquero, Negro y Otro. |
| NOMBRE_DEPARTAMENTO | Texto | Corresponde al nombre del departamento de residencial del reportado positivo por COVID-19. |
| NOMBRE_MUNICIPIO | Texto | Corresponde al nombre del municipio de residencia del reportado positivo por COVID-19. |

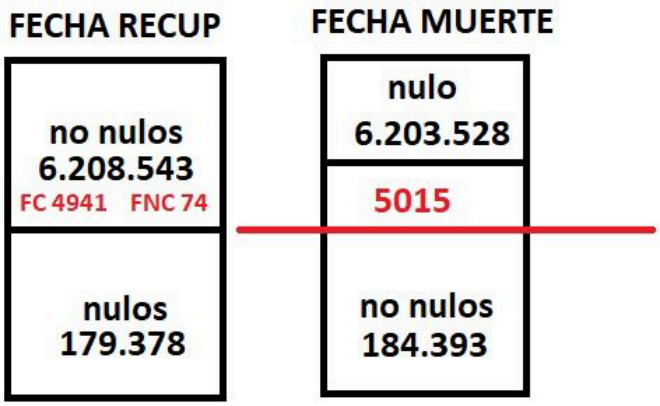
# Fase 4: exploración de datos

### Variables de caracterización de la población




- Al examinar las variables **ubicación del caso, estado y recuperado,** observamos que las tres categorías principales (*casa, leve y recuperado*) presentan la **misma frecuencia**. Además, las frecuencias de las demás categorías son bastante similares, lo que sugiere una posible relación entre ellas.

Sin embargo, al comparar la suma de las frecuencias de todas las categorías con la cantidad total de datos en la base de datos (6.387.921), se detecta una discrepancia de 5.015 casos. Esta diferencia nos llevó a investigar la situación en detalle, llegando a la siguiente imagen:



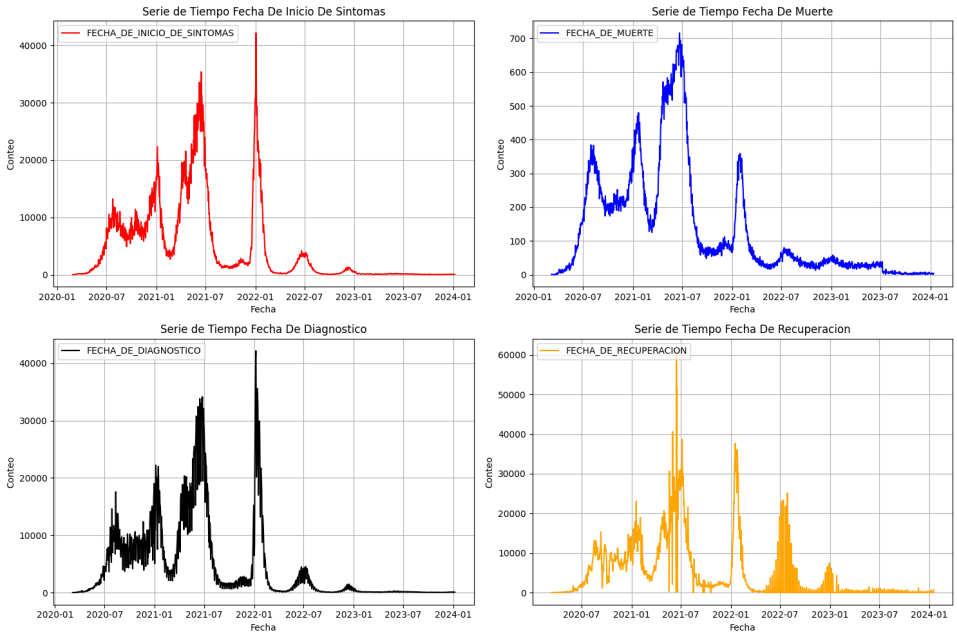

La investigación detallada reveló que los 5.015 casos que generaron la discrepancia corresponden a pacientes que inicialmente fueron clasificados como *recuperados* del **COVID-19**, pero que posteriormente *fallecieron* a causa de la enfermedad. Esta situación se explica por el hecho de que el sistema de categorización no siempre se actualiza de manera inmediata, lo que puede generar un desfase temporal entre la recuperación inicial y el posterior fallecimiento.

- Al examinar la variable **unidad de medida de edad**, observamos que en el 99,4% de los casos registrados la categoría es *años*. Esto indica que la población afectada por el **COVID-19** en el conjunto de datos está compuesta principalmente por personas de **al menos un año de edad**. Esta información es relevante evaluar el impacto del virus en diferentes segmentos de la población.

- Al examinar la variable **sexo**, observamos que las frecuencias de las categorías *masculino* y *femenino* son bastante similares. Sin embargo, la categoría *femenino* presenta una frecuencia ligeramente mayor, lo que sugiere una posible diferencia en la distribución de casos por sexo.

- Al examinar la variable **pertenencia étnica**, observamos que el 96,5% de los casos ha utilizado la categoría *otros* como identificación de su etnia. Esta alta proporción sugiere que una gran parte de la población afectada por **COVID-19** en el conjunto de datos no se identifica con ningún grupo étnico reconocido o representa un caso muy especial que no se ajusta a las categorías predefinidas.

## Variables temporales



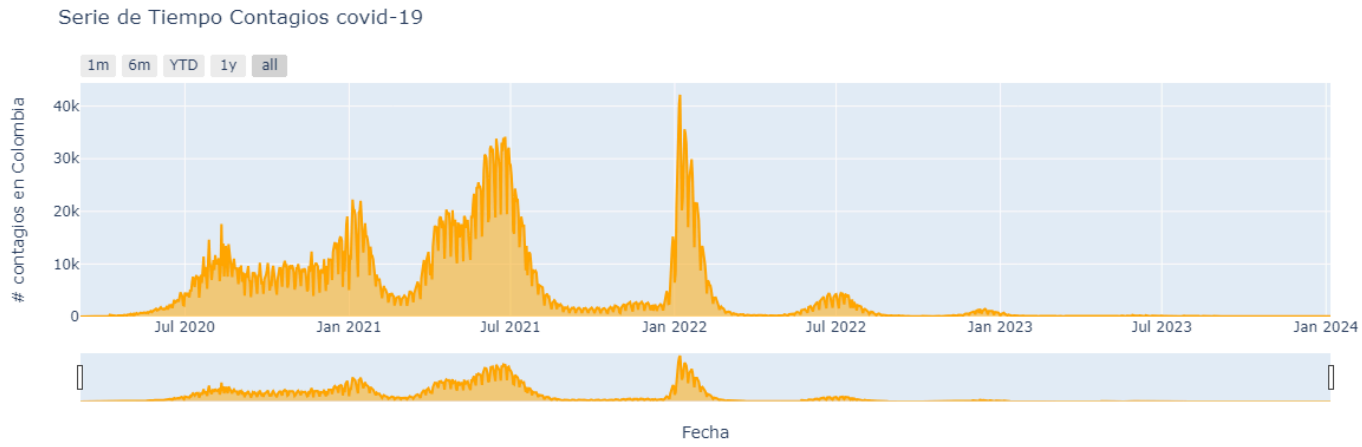

En general, las cuatro gráficas correspondientes a los períodos comprendidos entre *enero de 2020* y *enero de 2022* presentan un comportamiento similar. Además, las escalas utilizadas para las variables **fecha de diagnóstico, fecha de inicio de síntomas y fecha de recuperación** son comparables. Sin embargo, en el caso de la variable **fecha de muerte**, la escala empleada es significativamente diferente debido a la menor cantidad de datos disponibles.

- En cuanto a la **fecha de inicio de síntomas**, se observa un comportamiento con picos importantes en *fechas específicas*. Además, se detectan cambios de *tendencia* muy marcados en un *corto período de tiempo*, lo que coincide con la proximidad de las fechas de los picos. Cabe destacar que la naturaleza de esta variable, al ser un inicio de síntomas informado, implica que en varias ocasiones coincide con la **fecha de diagnóstico**.

- En cuanto a las **fechas de fallecimiento,** observamos un comportamiento similar al de las **fechas de diagnóstico**, con un leve **desplazamiento** de los picos. Esto podría explicarse por el hecho de que los picos de **alto contagio** precedían a los picos de *fallecimientos*, lo cual tiene sentido desde el punto de vista **epidemiológico**.

- En cuanto a la **fecha de recuperación**, se observan momentos claros en los que el valor desciende a 0, lo que corta la serie de manera **abrupta** en varios puntos. A partir de fechas posteriores a *mayo de 2022*, aproximadamente, la serie se asemeja más a un histograma con barras puntuales. Debido a su complejidad, esta serie no se incluirá en el tablero.

- Para la fecha de diagnóstico, realizaremos un análisis más detallado utilizando una perspectiva más amplia




La variable **fecha de diagnóstico** presenta características distintivas en comparación con la variable **fecha de inicio de síntomas**. Si bien la curva de diagnóstico es menos *suave* que la de inicio de síntomas, la naturaleza de la información de diagnóstico permite corroborar de manera definitiva un caso positivo de **COVID-19**. Esta característica la convierte en una variable de gran calidad para analizar el comportamiento de los casos confirmados (aspecto que se considera para su integración en el Business Intelligence).

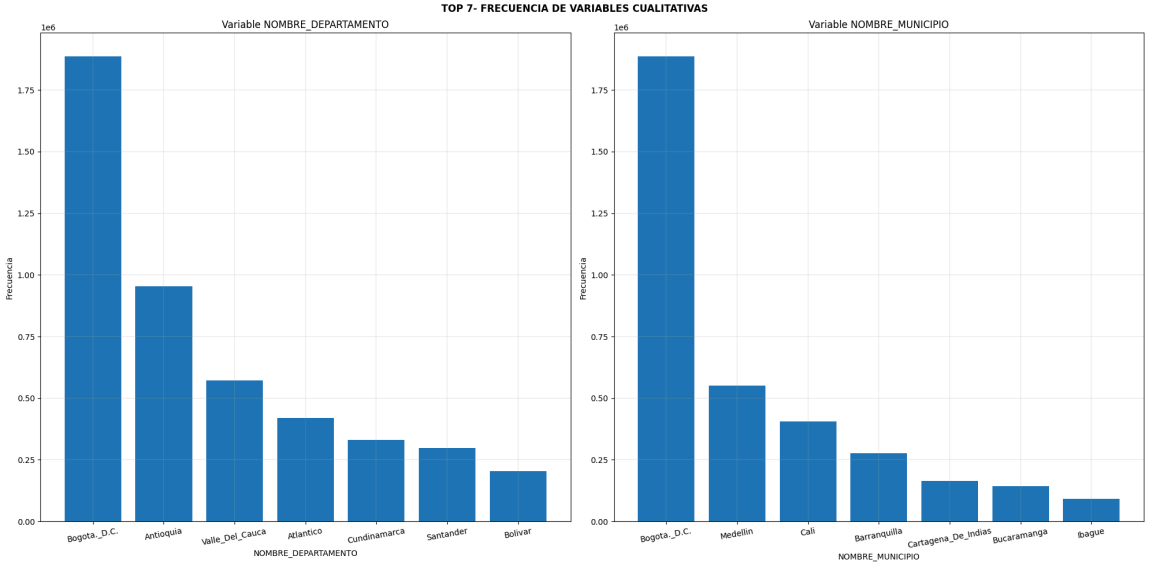
- *Varianza marginal fluctuante*: La serie temporal de la variable **fecha de diagnóstico** presenta una varianza marginal fluctuante, lo que significa que el rango de valores de la variable se **amplia o disminuye a medida que transcurre el tiempo.**

- *Tendencia cambiante*: La tendencia de la variable fecha de diagnóstico fluctúa rápidamente y, en ocasiones, lo hace de manera **abrupta** con grandes saltos o caídas. Este comportamiento evidencia una volatilidad considerable y un patrón similar a una caminata aleatoria.

- *Potencial estacionalidad*: Se observa un comportamiento que sugiere la existencia de estacionalidad en la variable **fecha de diagnóstico**. Sin embargo, la frecuencia de este patrón estacional también presenta fluctuaciones a lo largo del tiempo.

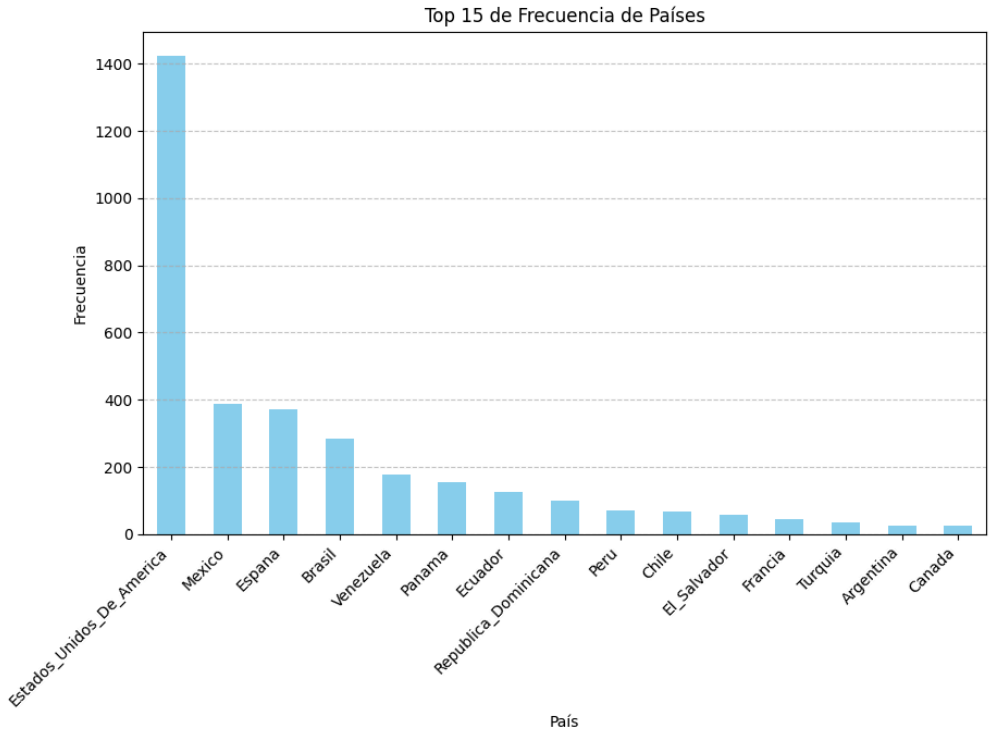
- *Análisis por periodos*: Debido al amplio rango de valores que puede tomar la variable **fecha de diagnóstico**, se observa que para las fechas del 2023 al 2024 la curva se comporta como una línea recta. Este comportamiento podría requerir un análisis aparte más enfocado en estos periodos específicos.

## Variables espaciales




- Observamos que la frecuencia de casos en *Bogotá* es casi el doble que en *Antioquia*. Además, este top 7 de departamentos con mayor frecuencia de casos coincide con los departamentos con mayor población del país.
- **Bogotá** presenta una frecuencia de casos más de *tres veces superior a la de Medellín*, que ocupa el segundo lugar. En cinco de los siete municipios del top 7, la capital del departamento coincide con la posición del departamento en su respectivo ranking. Esto indica que el comportamiento de estos municipios está bien reflejado por el comportamiento a nivel departamental. (tendremos esto en cuenta para el BI).

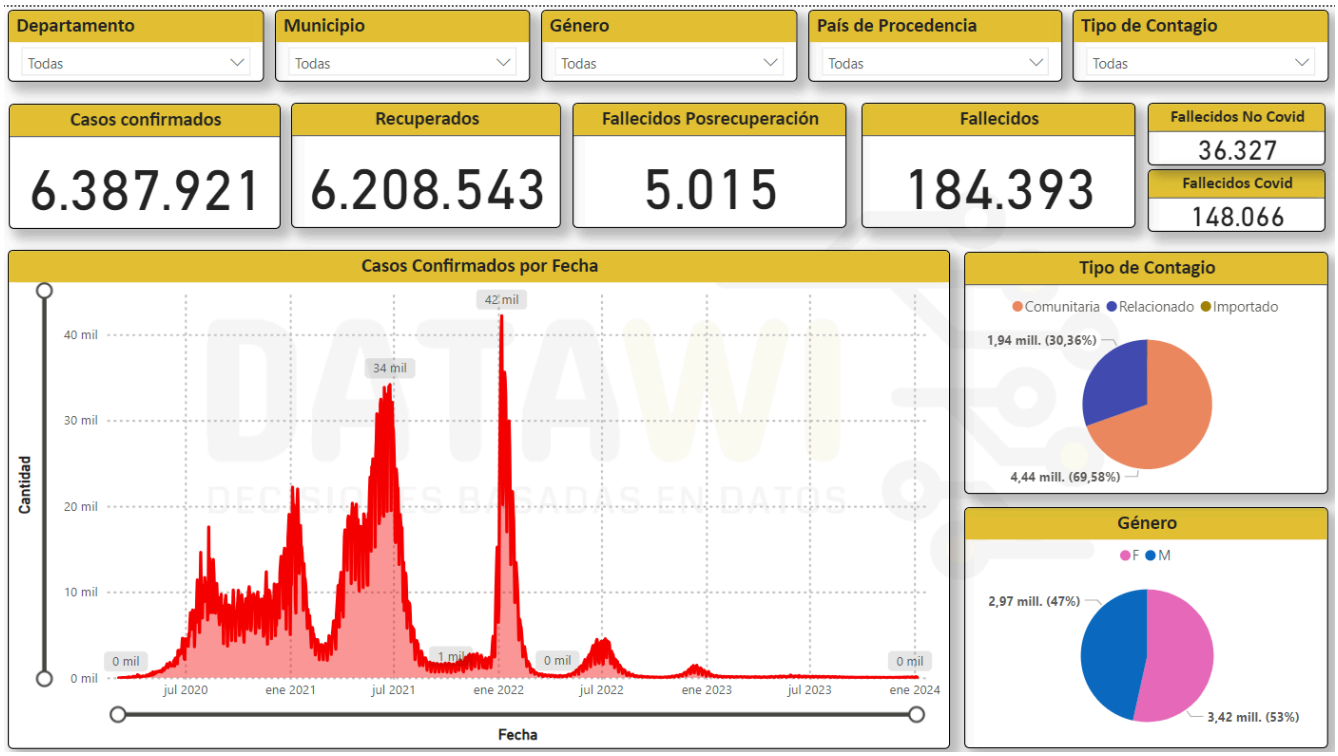
La columna denominada **Nombre del país** contiene información sobre el país de posible contagio del virus para cada individuo. Esta variable para poder ser analizada primero será filtrada sin **Colombia**.




Queda claro que **Estados Unidos de América**, con un número de casos tres veces mayor que el de *México*, que ocupa el segundo lugar en el ranking, es el claro dominador. Además, en el top 15, nueve de los países son *hispanohablantes*. Cabe destacar que solo tres países de Europa completan el top 15, mientras que el resto pertenece al continente americano. Es importante mencionar que en este top 15 se concentra el 90.3% aproximadamente de los casos registrados con procedencia diferente a **Colombia**.

# Fase 5: evaluación y despliegue

## Tablero con datos específicos:



# Tablero con datos sobre mapa