**RETO # 2 – CALIDAD DE DATOS**

**Sergio Alejandro Peña Pinto**

**Andrés David Ruiz Fierro**

**Daniela Ruiz Valencia**

A. Plantear preguntas de negocio que se puedan resolver a partir de los datos

B. A partir de las preguntas de negocio, analizar la calidad de los datos para poder responder las preguntas de negocio

C. Reglas que se puedan aplicar para la gestión de calidad para los datos.

**A). PREGUNTAS DE NEGOCIO:**

1. Desde el inicio de la pandemia por el COVID-19, ¿en qué fecha se presentaron más casos positivos?
2. Desde el inicio de la pandemia por el COVID-19, ¿en qué mes se presentaron más casos positivos?
3. Desde el inicio de la pandemia por el COVID-19, ¿cuál departamento ha tenido más casos reportados?
4. Desde el inicio de la pandemia por el COVID-19, ¿cuál municipio ha tenido más casos reportados? (Excepto Bogotá)
5. En un mes específico, ¿cuál departamento tuvo más casos positivos de COVID-19?
6. En un mes específico, ¿cuál municipio tuvo más casos positivos de COVID-19?
7. ¿Qué rangos de edad son los más afectados por el COVID-19?
8. ¿Es más propenso a contagiarse del COVID-19 el género masculino o femenino?
9. ¿En qué mes se ha reportado la mayor cantidad de fallecimientos por COVID-19 en el país?
10. ¿En qué día de cada mes se reportó la mayor cantidad de fallecimientos por COVID-19 en el país?
11. Desde el inicio de la pandemia por el COVID-19, ¿en qué departamento se presentaron más fallecimientos?
12. ¿Cuáles son los municipios con más reportes de fallecimientos por el COVID-19 en cada departamento del país?
13. Desde el inicio de la pandemia por el COVID-19, ¿de qué países han llegado más casos positivos a Colombia?
14. En lo transcurrido de la pandemia por el COVID-19, ¿cuál ha sido el tipo de contagio más frecuente?
15. ¿Cuál es el tiempo promedio de recuperación de las personas contagiadas por el COVID-19?
16. En promedio, ¿cuánto tiempo tardan las personas en notificar el posible contagio por COVID-19, desde la fecha de inicio de los síntomas?
17. Teniendo en cuenta las personas que han fallecido por COVID-19, ¿cuál es el tiempo promedio que transcurre desde la fecha de diagnóstico a la fecha de fallecimiento?
18. Desde el inicio de la pandemia por el COVID-19, ¿cuáles son los grupos étnicos que han resultado más afectados?
19. ¿En qué mes se ha reportado la mayor cantidad de recuperados del COVID-19 en el país?

**B). HALLAZGOS DE LOS DATOS:**

A continuación, se analizará la calidad de los datos por cada columna del set de datos.

**Fecha reporte Web, Fecha de notificación, Fecha de inicio de síntomas, Fecha de diagnóstico, Fecha de recuperación**

Se presenta un formato compuesto por la fecha y la hora en la que se realizó el reporte, sin embargo, la hora registrada siempre marca las 12:00:00 AM, lo cual es un dato que no genera beneficio ni utilidad a la base de datos. En lo referente a la fecha únicamente (día, mes, año), todas tienen el mismo formato (DD/MM/AAAA).

**ID de caso**

El dato que almacena es un valor numérico único de carácter auto incremental que posee cada uno de los casos registrados en la base de datos. Por esta razón, este dato no presenta alguna irregularidad que no permita dar respuesta a las preguntas planteadas.

**Código DIVIPOLA departamental y municipal**

El dato que almacena es un valor numérico que sirve como identificador para un departamento y un municipio. Dicho identificador es un dato cruzado de las bases de datos del DANE y no posee ninguna inconsistencia que no permita dar respuesta a las preguntas planteadas.

**Nombre departamento y municipio**

El dato que se almacenan en estas columnas es de tipo VARCHAR, y guarda el nombre del departamento y el municipio para el caso registrado. Este dato es fundamental para dar respuesta a las preguntas planteadas ya que en varias se pide reportar el nombre del departamento o municipio que cumpla con las características buscadas en las preguntas.

**Edad**

La columna almacena un valor numérico que representa la edad de la persona registrada en la base de datos. Este dato es de utilidad para dar respuesta a preguntas que estén relacionadas con la edad de las personas positivas para COVID-19.

**Unidad de medida de edad**

La columna almacena un valor numérico del 1 al 3, donde en la mayoría de los casos es 1. Sería necesario indagar con el cliente el significado de este valor, para así saber en qué podría ser útil.

**Sexo**

La columna almacena un valor de tipo VARCHAR que representa el sexo de la persona registrada en el sistema. Este dato no presenta ninguna irregularidad y es fundamental para dar respuestas a preguntas relacionadas con el sexo de las personas y otros factores.

**Tipo de Contagio**

La columna almacena un valor de tipo VARCHAR que muestra el tipo de Contagio de la persona registrada. Este dato no presenta ningún tipo de error, pero si sirve para dar respuesta a parte de las preguntas planteadas.

**Ubicación del caso**

Se almacena un valor de tipo VARCHAR que representa la ubicación en que se encuentra a la fecha cada uno de los casos registrados (Leve, Moderado, Grave o Fallecido). Sin embargo, se recomienda eliminar la opción de Fallecido con el fin de que esta opción solo la maneje la columna “Recuperado”, ya que es importante conocer el historial de una persona que ya ha fallecido.

**Estado**

Se almacena un valor de tipo VARCHAR que muestra el estado de la persona registrada a la fecha. Al igual que en la columna “Ubicación del caso” se debe eliminar la opción Fallecido por la misma razón expuesta en la descripción de la anterior columna.

**Código ISO del País**

Se almacena un valor de tipo numérico que sirve como identificador para un determinado país. Dicho identificador es un dato cruzado del estándar internacional de normalización ISO 3166-1. En esta base de datos no se muestra el identificador de Colombia, por lo cual sólo nos deja ver NULL en estas filas.

**Nombre del País**

Se almacena un valor de tipo VARCHAR que representa el nombre del país del cual fue importado el contagio por COVID-19. Para este caso, no se debe realizar uso de caracteres especiales, ni se deben dejar espacios en blanco NULL para los casos que no fueron importados de otro país.

**Recuperado**

Se almacena un valor de tipo VARCHAR que determina el estado en que se encuentra el caso de COVID-19 registrado manejando 4 opciones principales: Fallecido, Recuperado, Activo y N/A. Este dato no presenta ningún hallazgo de error, y es fundamental para realizar los reportes históricos acerca de los casos de COVID-19 registrados en la base de datos.

**Tipo de recuperación**

Se almacena un valor de tipo VARCHAR que presenta como se reportó la recuperación de la persona registrada, manejando tres valores principales: PCR, Tiempo, y un campo vacío NULL. Para este caso se observó la presencia de espacios en blanco NULL los cuales deben cambiarse por otra opción.

**Pertenencia étnica**

Esta columna almacena un valor de tipo numérico, que va desde el 1 al 6 (excepto el número 4). Sería necesario indagar con el cliente el significado de este código. Por el momento, esta columna no aporta valor para responder las preguntas de negocio planteadas en el numeral anterior.

**Nombre del grupo étnico**

Esta columna almacena un valor de tipo VARCHAR y representa el nombre del grupo étnico al que pertenece la persona del caso registrado. Dichos datos tienen valores que están en blanco lo cual al visualizar nos muestra NULL en esta columna y también podemos encontrar caracteres especiales.

**C). REGLAS DE CALIDAD DE DATOS:**

**Fecha reporte Web**

Esta columna de datos debe ser de tipo *Fecha:* DD/MM/AAAA. No es necesario que sea de tipo *Fecha y hora*, ya que la hora siempre es la misma (12:00:00 a.m.), y esto no aporta ningún valor para el análisis de la información. Adicionalmente, es necesario verificar que las fechas estén en orden ascendente.

**Fecha de notificación, Fecha de inicio de síntomas, Fecha de diagnóstico, Fecha de recuperación**

Esta columna de datos debe ser de tipo *Fecha:* DD/MM/AAAA. No es necesario que sea de tipo *Fecha y hora*, ya que la hora siempre es la misma (12:00:00 a.m.), y esto no aporta ningún valor para el análisis de la información.

**ID de caso**

Esta columna de datos debe ser de tipo *Número entero.* Se debe asegurar que el último registro sea igual al número total de filas de la base de datos.

**Código DIVIPOLA departamental y municipal**

Esta columna de datos debe ser de tip*o Número entero.* Se debe asegurar que los datos consignados correspondan a los que se muestran en la base de datos del DANE (<https://geoportal.dane.gov.co/geovisores/territorio/consulta-divipola-division-politico-administrativa-de-colombia/>).

**Nombre departamento y municipio**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan caracteres especiales y todos los registros sean pertenecientes al país.

**Edad**

Esta columna de datos debe ser de tipo *Número entero.* Se debe asegurar que no se presenten datos atípicos que correspondan a edades muy mayores, esto se podría validar con datos cruzados de la Registraduría Nacional de la Nación.

**Unidad de medida de edad**

Esta columna de datos debe ser de tipo *Número entero.* Por el momento el significado no tiene algún valor válido, por lo cual no será tenido en cuenta en el análisis de las preguntas de negocio.

**Sexo**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que solo existan datos con valor F o M en mayúsculas y que no existan campos nulos.

**Tipo de Contagio**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan campos nulos, ni con caracteres especiales. Además, los datos deben estar entre las siguientes categorías: Importado, Relacionado o En estudio.

**Ubicación del caso**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan campos nulos, ni con caracteres especiales. Además, es recomendable eliminar la opción “Fallecido” de las opciones que posee este campo, la cual solo estará en la columna llamada “Recuperado”.

**Estado**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan campos nulos, ni caracteres especiales. Además, se debe eliminar la opción “Fallecido” de las opciones que posee este campo, la cual solo estará en la columna de “Recuperado”.

**Código ISO del País**

Esta columna de datos debe ser de tipo *Número entero.* Se debe asegurar que los campos NULL se rellenen con el número 0 y que los datos consignados correspondan a los que se muestran en el estándar internacional de normalización ISO 3166-1, el cual se puede visualizar en (<http://utils.mucattu.com/iso_3166-1.html>).

**Nombre del País**

Esta columna de datos debe ser de tipo *Texto.* Igualmente se debe asegurar que los campos no sean nulos, ni posean caracteres especiales (tildes o letras ñ). Por ejemplo, el país ESPAÑA aparece de la siguiente forma “ESPAÃ’A”, por tanto, para que se guarde de la manera correcta debería ser “ESPANA".

**Recuperado**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan caracteres especiales y se utilicen únicamente las opciones anteriormente mencionadas (Fallecido, Activo, N/A, Recuperado).

**Tipo de recuperación**

Esta columna de datos debe ser de tipo *Texto.* Se debe asegurar que no existan caracteres especiales y se utilicen únicamente las opciones anteriormente mencionadas (PCR, Tiempo) y a los espacios que estén vacíos, asignarles N/A.

**Pertenencia étnica**

Esta columna de datos debe ser de tipo *Número Entero.* Sin embargo, se determinó que no aporta valor para responder las preguntas planteadas y, además, no se sabe cuál es su uso específico.

**Nombre del grupo étnico**

Esta columna de datos debe ser de tipo *Texto.*  De igual manera que en los anteriores datos de tipo texto, se debe evitar dejar espacios vacíos o textos con caracteres especiales. Para el caso de espacios vacíos, preferiblemente se les debe asignar N/A.

**Resumen:**

|  |  |  |
| --- | --- | --- |
| **TEXTO** | **NÚMERO ENTERO** | **FECHA** |
| No espacios nulos.  No caracteres especiales.  Asignar N/A a espacios vacíos. | No espacios nulos.  Asignar 0 a espacios vacíos.  Tener en cuenta el formato de cada columna que posea números enteros. | Quitar la hora y dejar la fecha en formato DD/MM/AAAA.  Tener en cuenta el formato de cada columna que posea fechas. |