

ANÁLISIS COVID-19 EN MÉXICO

PRESENTACIÓN DE
RESULTADOS



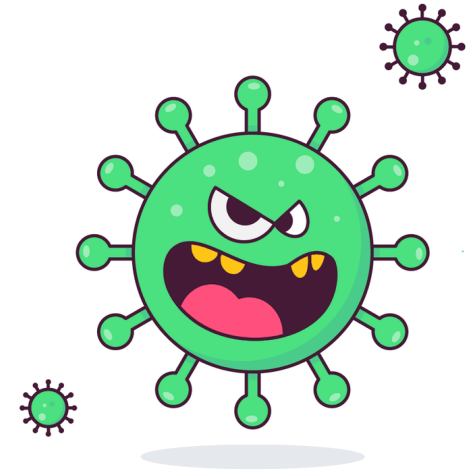


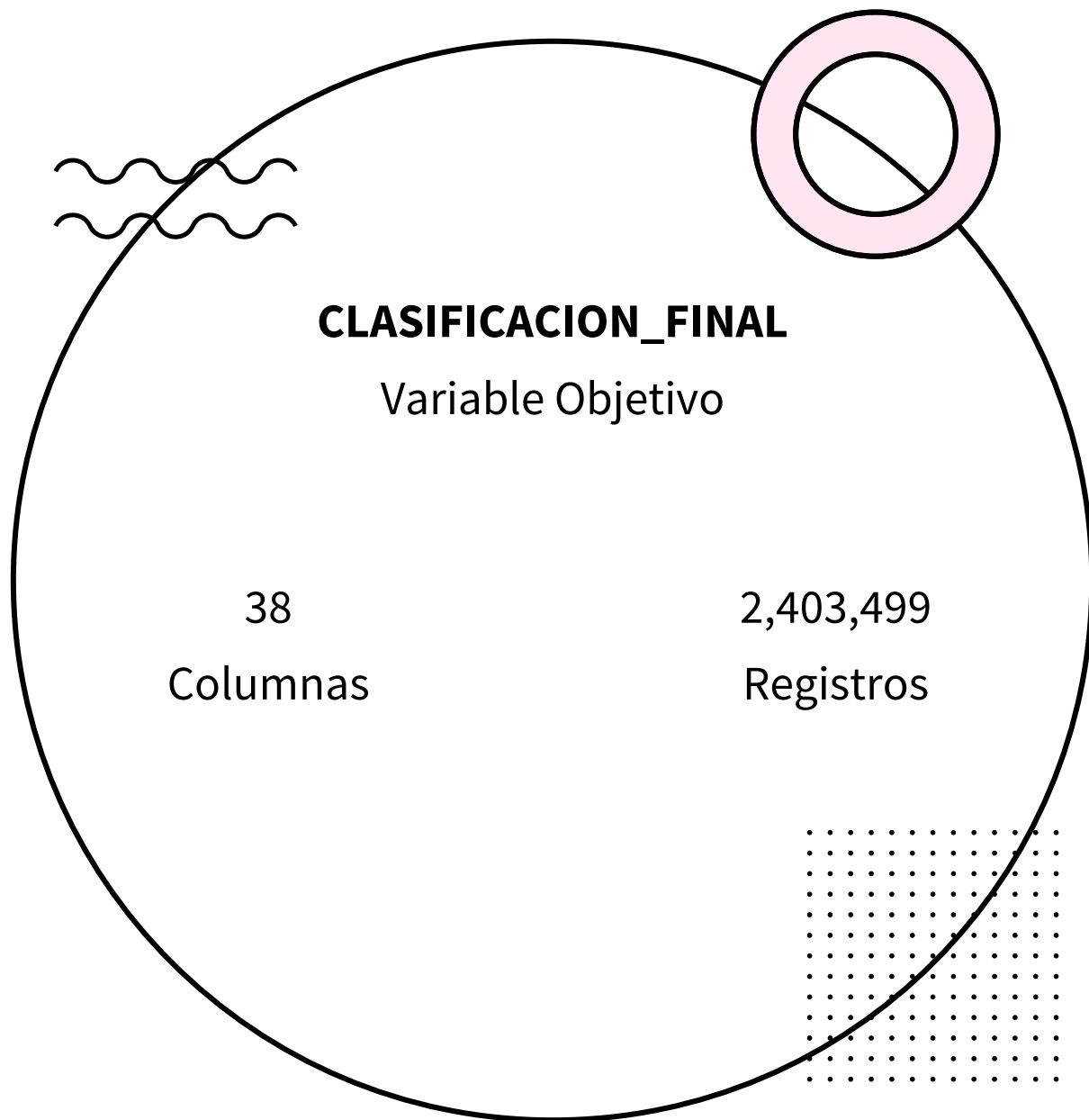
Objetivo

Crear una tabla con la información necesaria y preparada para desarrollar un modelo que sea capaz de predecir la clasificación final de un paciente en función de su edad, padecimientos médicos, lugar de residencia, entre otros. Se clasificará como negativo (1, 0), sospechoso (0, 1) o confirmado (0, 0).

Beneficios:

- Obtención de información relevante.
- Estructurar los datos de la manera más adecuada.





Conjunto de datos

Tabla que contiene los registros diarios de pacientes que fueron atendidos por **COVID-19** en México desde el 1º de enero del 2020 hasta el 31 de octubre del 2020.

Cuenta con contenido desagregado por sexo, edad, nacionalidad, padecimientos asociados, entre otros.

Calidad de datos



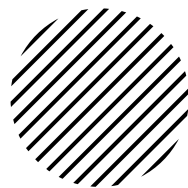
Datos no válidos:

Variables	Dato
v_SECTOR, v_ENTIDAD_NAC, v_INTUBADO, v_NEUMONIA, v_MIGRANTE	“no especificado”
v_EMBARAZO, v_DIABETES, v_EPOC, v_ASMA, v_HIPERTENSION, v_OBESIDAD	“se ignora”
v_PAIS_ORIGEN, v_PAIS_NACIONALIDAD	“se desconoce”

22
Variables

0%
Datos
duplicados

Calidad de datos



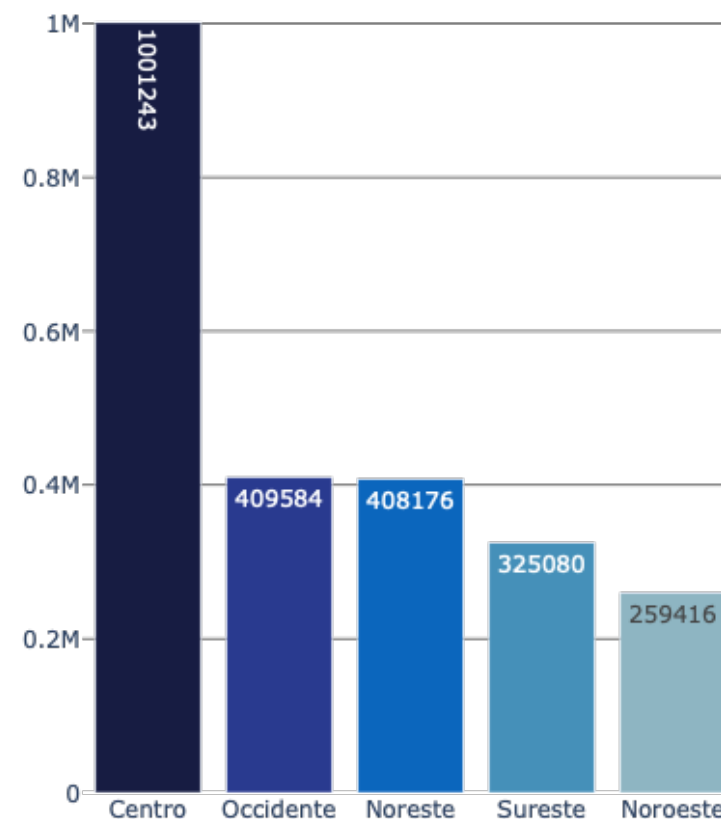
Variable	Compleitud
v_MIGRANTE	0.397%
v_OTRO_CASO	87.941%
v_HABLA LENGUA INDIG	96.287%
v_NEUMONIA	99.238%
v_ENTIDAD_NAC	99.525%
v_EMBARAZO	99.617%
v_DIABETES	99.663%
v_TABAQUISMO	99.673%
v_HIPERTENSION	99.684%

○ Análisis exploratorio

- El **42%** de los pacientes atendidos por COVID-19 provienen de la región centro.
- El número de pacientes de la región occidente y noreste es equivalente al **17%** cada una.
- La región sureste representa el **13%** de los pacientes totales.
- La región noroeste representa el **11%**.



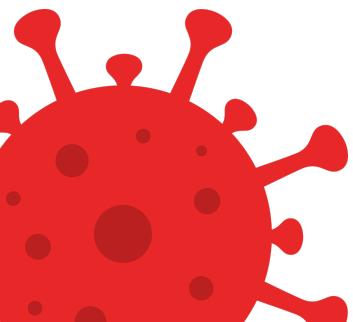
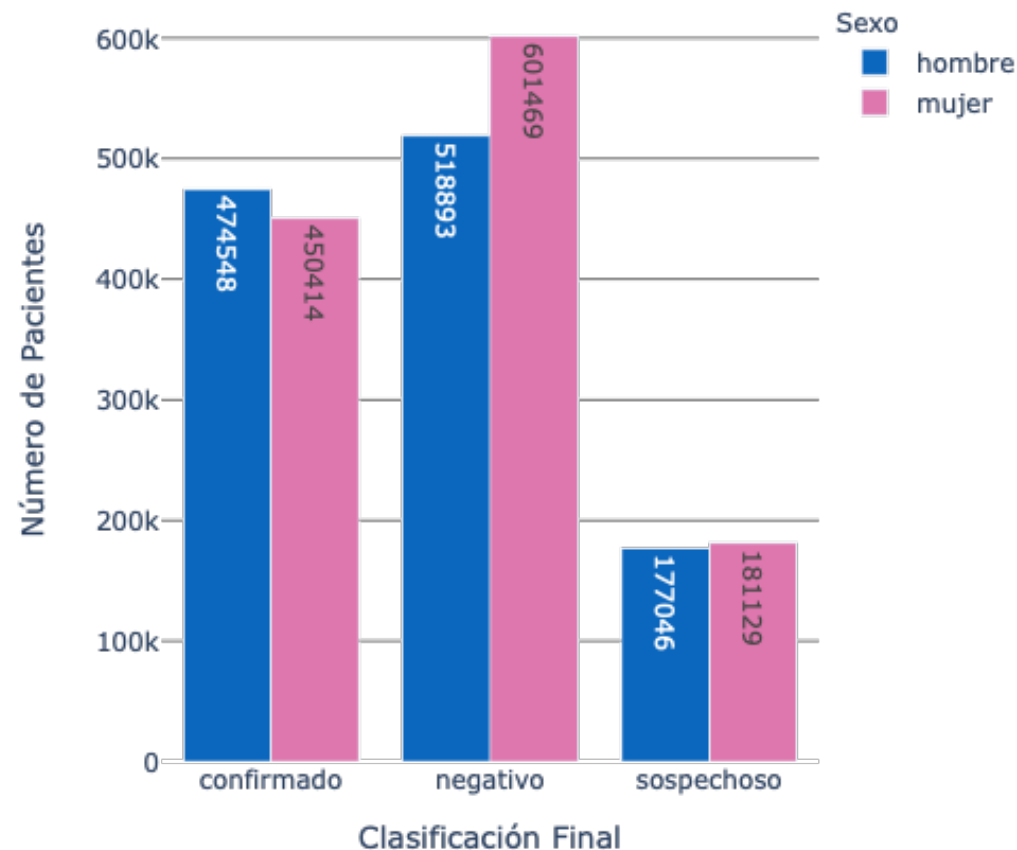
Número de Casos por Regiones en México



○ Análisis exploratorio

- De las pacientes el **36%** ha sido clasificada como caso confirmado, el **49%** como negativo y el **15%** como caso sospechoso.
- En cuanto a los pacientes el **40%** son casos confirmados, **44%** casos negativos y **16%** casos sospechosos.

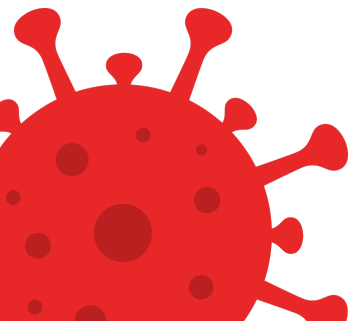
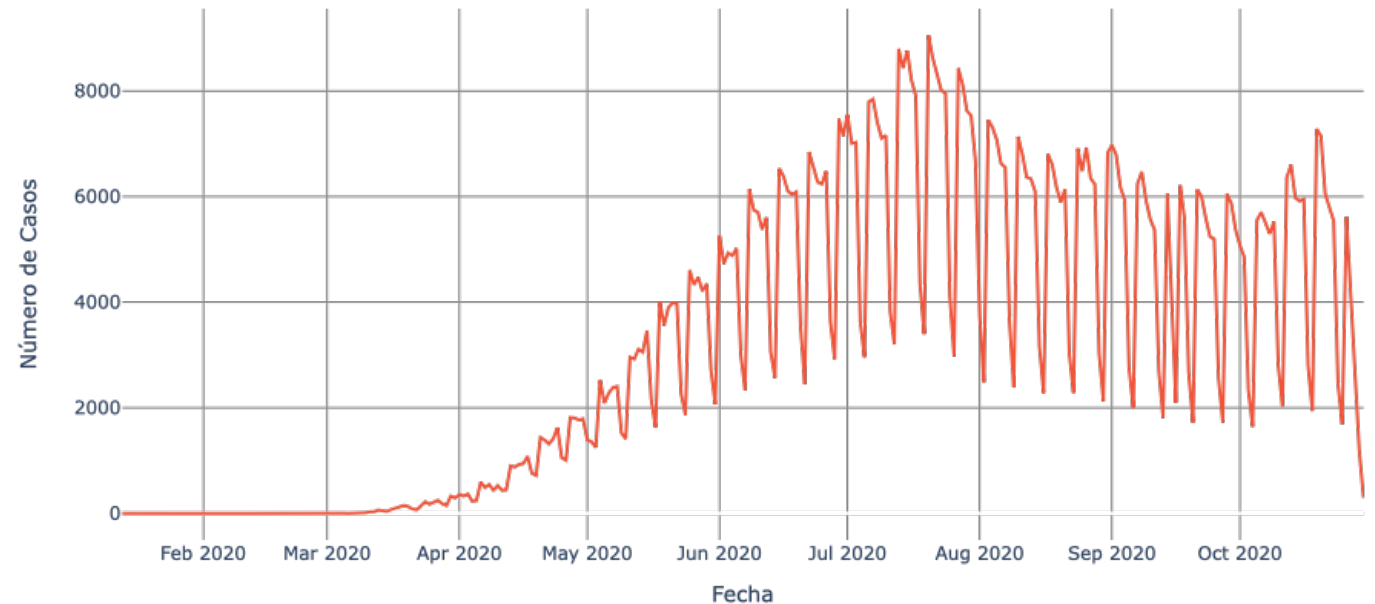
Total de Clasificaciones por Sexo



○ Análisis exploratorio

- El 20 de julio es el pico más alto con **9,051** casos confirmados.
- El 31 de octubre se alcanza el pico más bajo con **296** casos confirmados.
- En promedio al día se tienen **3,168** casos confirmados.

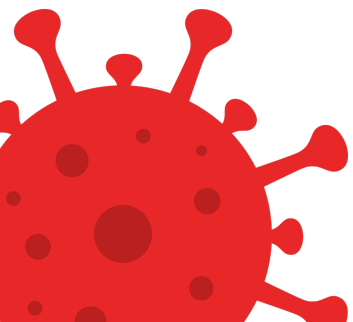
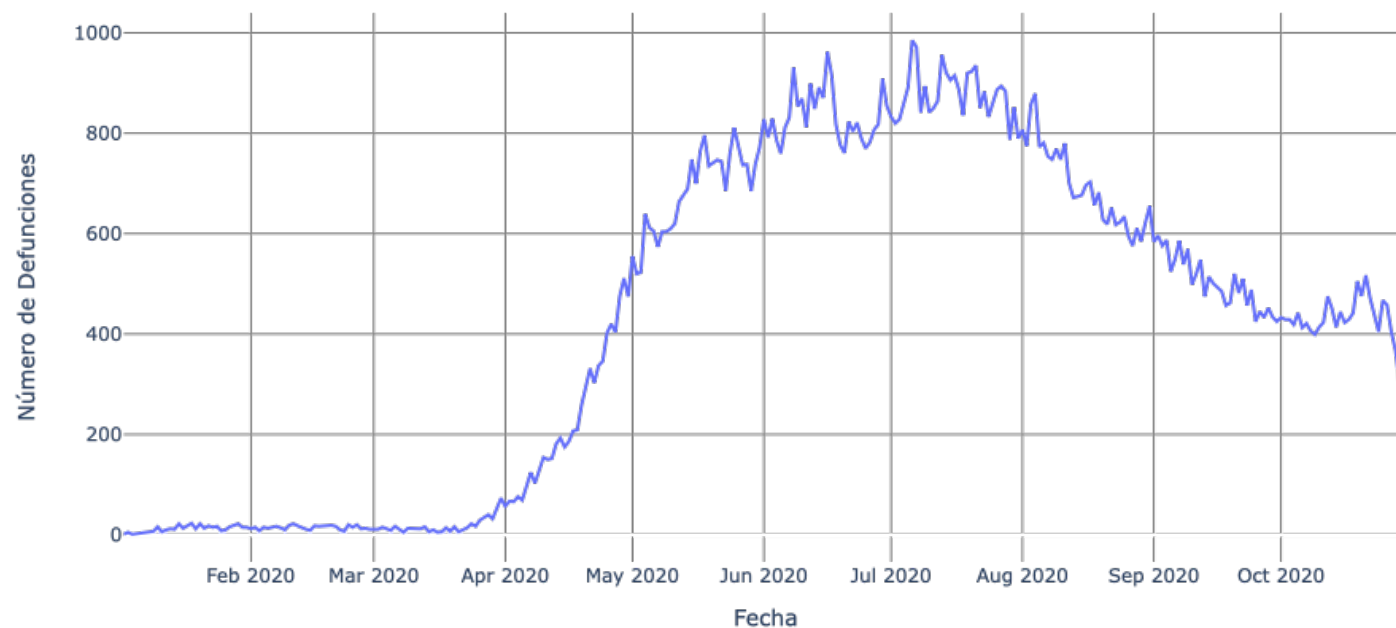
Número de Casos Confirmados Diario

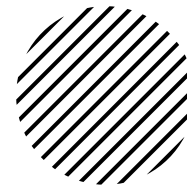


○ Análisis exploratorio

- El 6 de julio se alcanzó el pico más alto con **985** defunciones.
- El 31 de octubre se alcanza el pico más bajo con **1** defunción.
- En promedio al día se tienen **431** defunciones.

Número de Defunciones Diarias

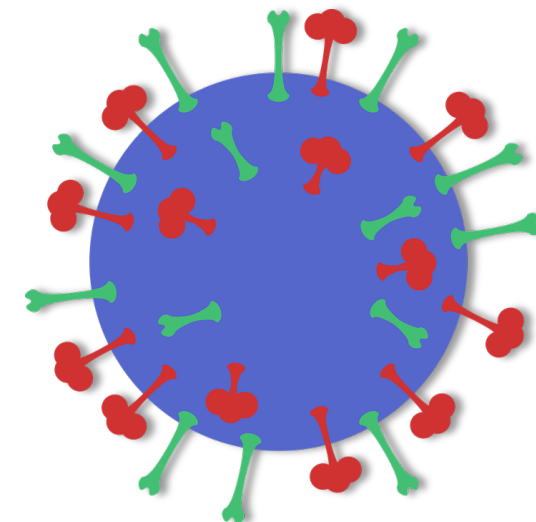




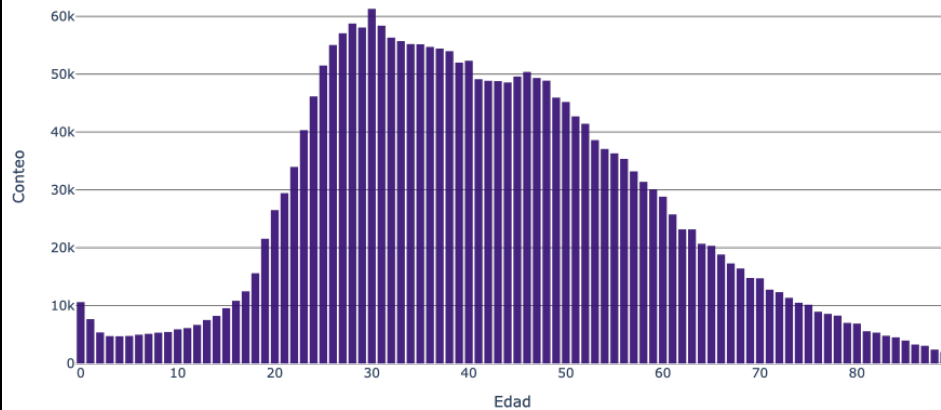
Datos anómalos

Dentro del mundo de variables que se tiene en el conjunto de datos, la única que era continua, presentaba datos atípicos.

- ✓ Se detectaron **7,684** datos atípicos.
- ✓ Se procedió a eliminarlos porque representaban un 0.32% del total.



Distribución de la Edad de los Pacientes






Datos faltantes

Como el conjunto de datos es una serie de tiempo que presenta tendencia.

- ❑ Los datos faltantes de las variables categóricas se imputaron mediante la **moda**.



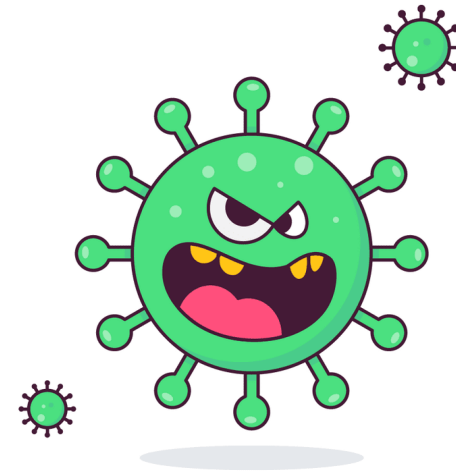
Variable	Valor imputado
c_EDAD	99.680%
v_OTRO_CASO	87.941%
v_HABLA LENGUA _INDIG	96.287%
v_NEUMONIA	99.238%
v_ENTIDAD_NAC	99.525%
v_EMBARAZO	99.617%
v_DIABETES	99.663%
v_TABAQUISMO	99.673%
v_HIPERTENSION	99.684%

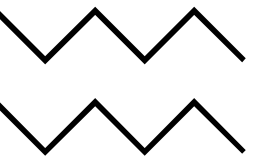




Ingeniería de variables

Dado que en su mayoría el conjunto de datos cuenta con variables de tipo categóricas, todas fueron transformadas a variables dummies, para poder tener una representación numéricas de estas, sin perder información.





Reducción de dimensiones

Tras haber creado variables dummies, el conjunto de datos aumentó significativamente en cuanto al número de variables, por lo que se eliminaron algunas tras aplicarles diferentes métodos que ayudaran a medir su desempeño en conjunto e individual.



- v_MIGRANTE
- v_ENTIDAD_UM
- v_ENTIDAD_NAC
- v_PAIS_ORIGEN
- v_MUNICIPIO_RES
- v_PAIS_NACIONALIDAD
- v_RESULTADO_LAB
- t_ID_REGISTRO

Relación de valor perdido

Filtro de alta correlación

Correlación (con el objetivo)

Multicolinealidad

- d_FECHA_ACTUALIZACION
- v_SECTOR_ssa
- v_EMBARAZO_no aplica
- v_INTUBADO_no aplica
- v_UCI_no aplica
- v_NEUMONIA_si
- v_HABLA LENGUA_INDIGENA_si
- v_NACIONALIDAD_mexicana
- v_TOMA_MUESTRA_si

51
Variables

17
Eliminadas



**¡GRACIAS
POR SU
ATENCIÓN!**