LTA Service Agreement UNDP México (UNDP-MEX-10029 / CSG-2024-002)

mottum. Analytica
Control y Gestión

# Análisis Exploratorio de Datos:

## Productores autorizados

## Programa del bienestar

Consultoría

"Diseño y programación de soluciones tecnológicas, ciencia de datos, análisis y visualización de datos"

SECRETARÍA EJECUTIVA SNA

29 de mayo 2024

PNUD

# ÍNDICE

# 1. Definición del problema.

## 1.1. Introducción

El programa de Fertilizantes es un programa a nivel federal, gestionado y coordinado a nivel estatal. Para llevar un mejor control del impacto de este programa de ayudas sociales, que consiste en el reparto de ayudas de fertilizantes a los productores, se publica en el portal Datos.gob.mx un listado de productores autorizados y un listado de productores beneficiarios. (Ver portal de datos abiertos).

El objetivo de la Secretaría Ejecutiva del Sistema Nacional de Anticorrupción (SESNA) es crear un padrón de beneficiarios que permita el análisis de este programa.

Este Análisis Exploratorio de Datos (EDA) hace referencia al Entregable 1 de esta consultoría: Listado de Productores Autorizados del Programa de Fertilizantes para el Bienestar 2023. El listado de beneficiarios se trabajará en los entregables 2 y 3.

| **Repositorio de código** |
| --- |
| https://github.com/MottumData/SESNA-Fertilizantes |

## 1.2. Definición del problema

El listado de productores autorizados del Programa de Fertilizantes para el Bienestar 2023. En su conjunto, corresponde a 44 bases de datos que contienen información de los productores autorizados a acceder al Programa de Fertilizantes para el Bienestar 2023 por entidad federativa y por número de convocatoria. El reto consiste en descargar y conciliar todas las bases de datos estandarizando los nombre de municipios y estados de cada productor autorizado.

Durante el proceso de recolección de información se identificaron múltiples errores manuales en el proceso de escritura de los municipios. Reconciliar dicho listado de 1.5 millones de filas requiere destreza a la hora de manejar los datos.

Como base de datos de referencia a la que conciliar, con los nombres de los estados y municipios se está tomando el Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Locales de INEGI.

## 1.3. Planteamiento de la solución

El enfoque propuesto consiste en estandarizar los nombres de los municipios y estados en ambas fuentes de datos quitando caracteres como guiones y mayúsculas, para luego hacer conciliar sobre nombres estandarizados dejando de manifiesto aquellos municipios sin emparejar.

Los municipios sin emparejar son aquellos que tienen un nombre realmente diferente en una base de datos que en otra. Para ello se ha calculado la distancia de Levenshtein, que calcula un ratio de la distancia vectorial entre dos strings. Esta técnica nos ha ayudado a reducir considerablemente el número de emparejamientos.

Hemos encontrado que sólo una porción pequeña de municipios (20 municipios aproximadamente) requieren revisión manual.

## 1.4. Notas importantes previas

**A.** Al intentar descargar las 44 bases de datos del portal de Datos.gob.mx, se ha detectado que sólo 25 han podido descargarse, mientras que las otras 19 enlaces llevan a páginas de error. Las últimas pruebas fueron realizadas a fecha 28 de mayo de 2024.

**B.** El proceso de reconciliación manual requiere intervención humana para emparejar aquellos municipios que no se ha identificado pareja con las técnicas antes mencionadas.

**C.** El listado de municipios sin emparejar se ha arreglado manualmente consultando fuentes de datos externas. Dentro de esta corrección manual cabe destacar que:

a. El municipio de ***Dr. Belisario Dominguez*** aparece en el estado de **Chiapas**. En la geografía política de México este nombre aparece en:

i. **Estado**: Chiapas | **Municipio**: Juárez | **Localidad**: *Dr. Belisario Dominguez*

ii. **Estado**: Chihuahua | **Municipio**: *Dr. Belisario Dominguez*

Se ha optado por la opción ii, modificar el estado a Chihuahua.

**D.** Además del municipio antes mencionado, se han modificado alrededor de 20 municipios más:

| ESTADO | MUNICIPIO |
|---|---|
| mexico | acambay |
| mexico | acambay |
| chiapas | belisario dominguez |
| chiapas | cintalapa |
| morelos | jonacatepec |
| oaxaca | san blas atempa |
| oaxaca | san juan mixtepec |
| oaxaca | san pedro totolapa |
| oaxaca | santiago chazumba |
| oaxaca | tezoatlan de segura y luna |
| morelos | tlaltizapan |
| jalisco | tlaquepaque |
| oaxaca | villa de tututepec de melchor ocampo |
| tlaxcala | yauhquemecan |
| oaxaca | zapotitlan del rio |
| tlaxcala | zitlaltepec de trinidad sanchez santos |
| veracruz de ignacio de la llave | Nan |

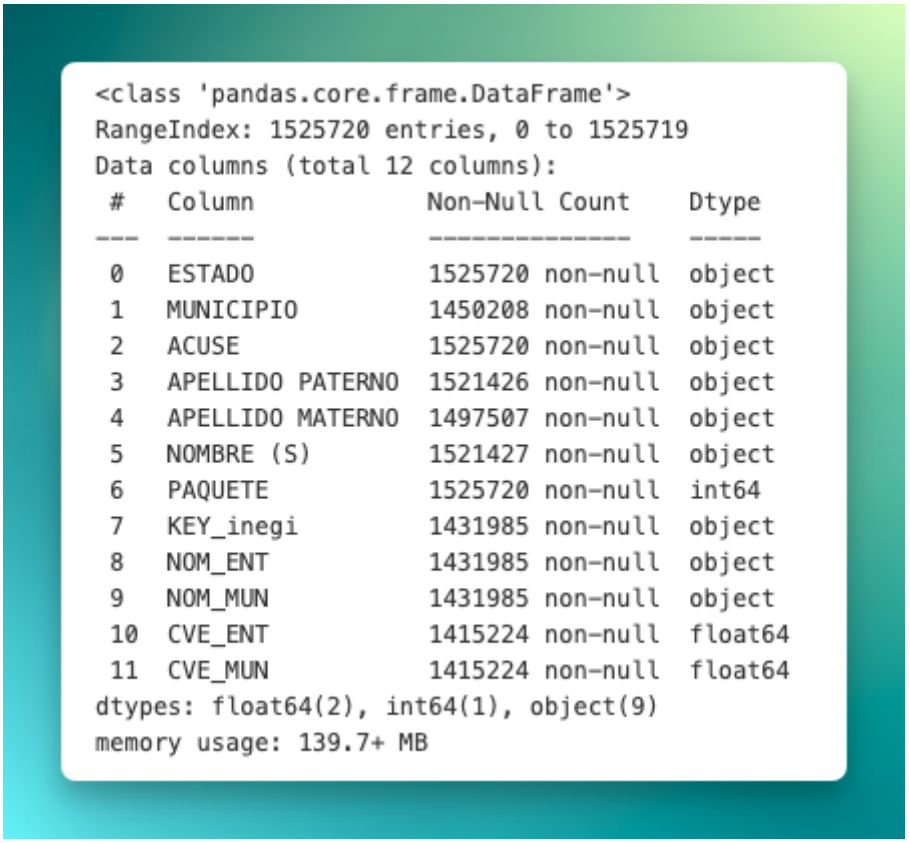
**E.** Los registros del ESTADO de **Veracruz de Ignacio de la Llave** no tienen MUNICIPIO asignado, por lo que aquellas secciones del análisis EDA relacionado con municipios, no tiene en cuenta los valores de este estado.

**F.** En análisis de datos se ha hecho sobre el dataset de Fertilizantes autorizados con los nombres de los municipios y sus códigos estandarizados según INEGI.

# 2. Análisis inicial.

## 2.1. Estructura de datos.

El dataset contiene 1.525.720 filas, cada una correspondiente a un productor autorizado.
El dataset final contiene 12 variables distribuidas de la siguiente manera.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525720 entries, 0 to 1525719
Data columns (total 12 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   ESTADO            1525720 non-null  object
 1   MUNICIPIO         1450208 non-null  object
 2   ACUSE             1525720 non-null  object
 3   APELLIDO PATERNO  1521426 non-null  object
 4   APELLIDO MATERNO  1497507 non-null  object
 5   NOMBRE (S)        1521427 non-null  object
 6   PAQUETE           1525720 non-null  int64
 7   KEY_inegi         1431985 non-null  object
 8   NOM_ENT           1431985 non-null  object
 9   NOM_MUN           1431985 non-null  object
 10  CVE_ENT           1415224 non-null  float64
 11  CVE_MUN           1415224 non-null  float64
dtypes: float64(2), int64(1), object(9)
memory usage: 139.7+ MB
```

**Nota importante:**

Para generar este dataset se ha pasado por un proceso previo de descarga de datasets, unión vertical de todos y unión horizontal con los código de Inegi. Por lo tando recordemos que las columnas de la 0 a la 6 pertenecen a la base de datos original mientras que de la 7 a la 11 son las generadas en el proceso previo. Dicho proceso previo de puede consultar en el python llamado *2_Data_cleaning_and_merged.ipynb.*

## 2.2. Información general

La información abarca una variedad de columnas que detallan desde datos personales hasta información de ubicación y de identificación. A continuación, se presenta un resumen general y algunas estadísticas descriptivas importantes.

- ESTADO y NOM_ENT: Ambas variables tienen un total de 20 y 19 valores únicos respectivamente, con GUERRERO siendo el más frecuente.

- MUNICIPIO y NUM_MUNi: Existen 1,694 y 1,674 municipios únicos respectivamente, con LAS MARGARITAS como el municipio más común.

- APELLIDO PATERNO, APELLIDO MATERNO, NOMBRE (S): Muestran la distribución de apellidos y nombres, siendo HERNANDEZ el apellido más común y JUAN el nombre más frecuente.

- PAQUETE: Tiene una distribución centrada en 1 y 2.

- CVE_MUN_Unique, CVE_ENT, CVE_MUN: Codificaciones que proporcionan una forma estándar de identificar entidades y municipios.

### 2.3. Estadísticas descriptivas.

| | index | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ESTADO | 1525720 | 20 | GUERRERO | 356223 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | MUNICIPIO | 1450208 | 1694 | LAS MARG... | 19497 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 2 | ACUSE | 1525720 | 1525720 | 23-PRONA... | 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | APELLIDO ... | 1521426 | 13551 | HERNAND... | 75301 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 4 | APELLIDO ... | 1497507 | 15970 | HERNAND... | 75248 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 5 | NOMBRE (S) | 1521427 | 116715 | JUAN | 24614 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 6 | PAQUETE | 1525720 | NaN | NaN | NaN | 1.4630993... | 0.498636... | 1 | 1 | 1 | 2 | 2 |
| 7 | KEY_inegi | 1431985 | 1713 | chiapas-la... | 19497 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 8 | NOM_ENT | 1431985 | 21 | Guerrero | 354841 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 9 | NOM_MUN | 1431985 | 1674 | Las Marga... | 19497 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 10 | CVE_ENT | 1415224 | NaN | NaN | NaN | 14.031049... | 5.862548... | 1 | 12 | 13 | 20 | 31 |
| 11 | CVE_MUN | 1415224 | NaN | NaN | NaN | 82.791531... | 104.45477... | 1 | 24 | 53 | 93 | 570 |

# 3. Limpieza de datos.

La limpieza de datos se ha llevado a cabo en dos partes. Primero nos hemos centrado en el dataset de inegi para luego pasar con el dataset de productores autorizados. Este proceso se puede seguir en el notebook 2_Data_cleaning_and_merged.ipynb.

### 3.1. Valores nulos.

La limpieza de valores nulos ha sido necesaria para el dataset de productores autorizados.

- Destacar que en el caso de Veracruz de Ignacio de la Llave nos encontramos que el dato de municipio para cada estado Veracruz de Ignacio de la Llave es siempre nulos en el conjunto de datos de productores autorizados. En total son 75.512
- Nombres ,apellidos paternos y apellidos maternos se han encontrado con valor de nulo. Con un total de 4.293 nombres, 4.294 apellidos paternos y 28.213 apellidos maternos.

## 3.2. Decisiones sobre valores nulos.

La limpieza de algunos valores nulos ha sido necesaria para el dataset de productores autorizados así como algunos requerimientos por parte de SESNA.

- En el caso de que el nombre y apellidos sean nulos (se muestran como NaN), y sí existe un número de acuse hemos sustituído estos valores por un campo de texto “unknown” (desconocido).

## 3.3. Valores duplicados.

No hemos encontrado valores duplicados en el conjunto de datos de Fertilizantes Autorizados.

## 3.4. Corrección de tipo de datos.

- Es importante asegurarse de que todas las variables excepto paquete son pasadas a string. CVE_MUN, CVE_ENT deben cumplir con este requisito para que no salgan identificados como un número entero y sí como las claves de municipios y entidades (los ‘0’ delante de cada clave se suprimen cuando es identificada como números enteros)

## 3.5. Formatear strings.

La función ‘clean_text’ usada en el Notebook ‘2_Data_cleaning_and_merged.ipynb’ es una utilidad diseñada para limpiar y estandarizar datos de texto para su posterior procesamiento o análisis.

- Primero, verifica si el texto de entrada está ausente o es nulo (NaN) y lo devuelve tal cual si es el caso.
- Posteriormente, elimina cualquier espacio en blanco al principio y al final del texto para asegurar que no haya espacios innecesarios al inicio o al final.
- La función convierte todos los caracteres a minúsculas, promoviendo la uniformidad y evitando problemas de sensibilidad a mayúsculas.
- También elimina los acentos de los caracteres utilizando normalizando los caracteres acentuados a sus formas básicas.
- Adicionalmente, la función utiliza una expresión regular para eliminar cualquier patrón específico que coincida con '- XXXX -', lo cual ayuda a limpiar segmentos de texto no deseados, encontrados en fase de EDA y posteriormente resuelto.
- Los múltiples espacios entre palabras son reemplazados por un solo espacio, asegurando un espaciado consistente en todo el texto.
- Finalmente, se eliminan los espacios restantes al principio y al final del texto.

Estos pasos, en conjunto, hacen que el texto sea consistente y limpio, mejorando la calidad y la fiabilidad de los datos para su uso posterior.

**Ejemplo:**

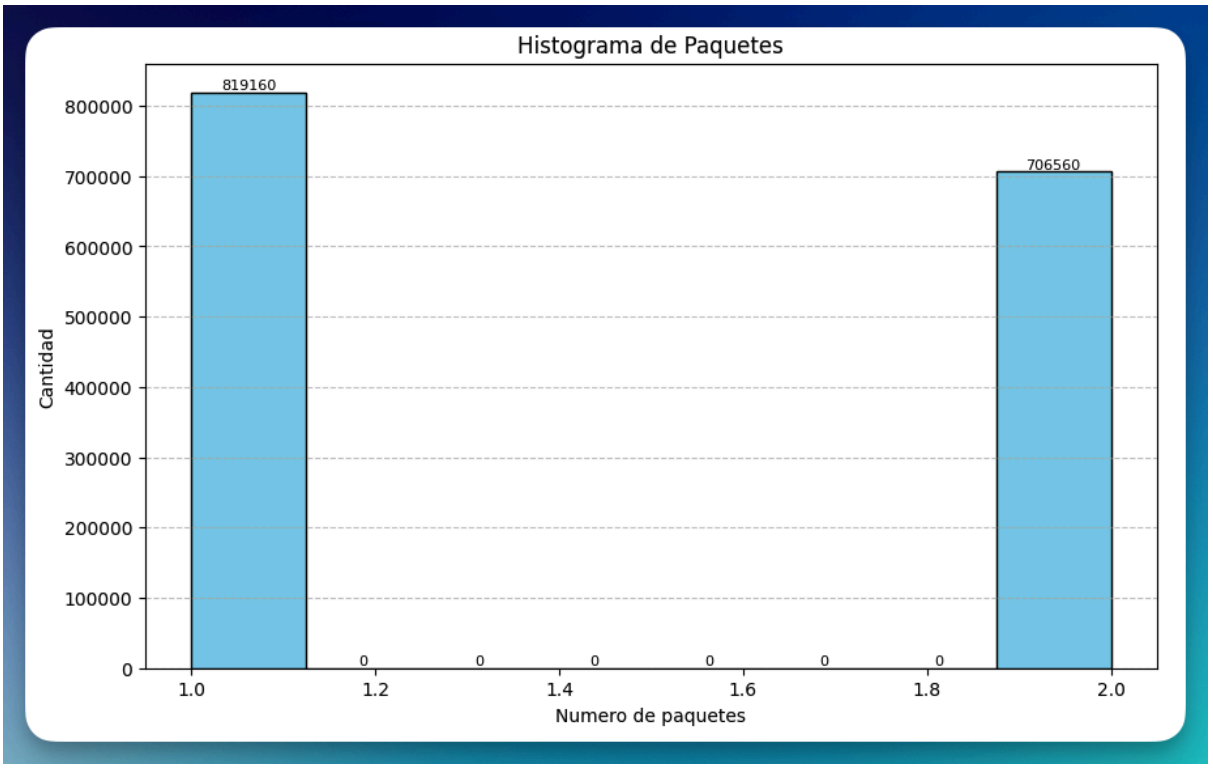
Antes: **"Pabellón de Arteaga"**

Después de limpieza: **"pabellon de arteaga"**

Esta funcionalidad es la que en gran parte nos ha permitido realizar el proceso de conciliación.

# 4. Análisis univariado.

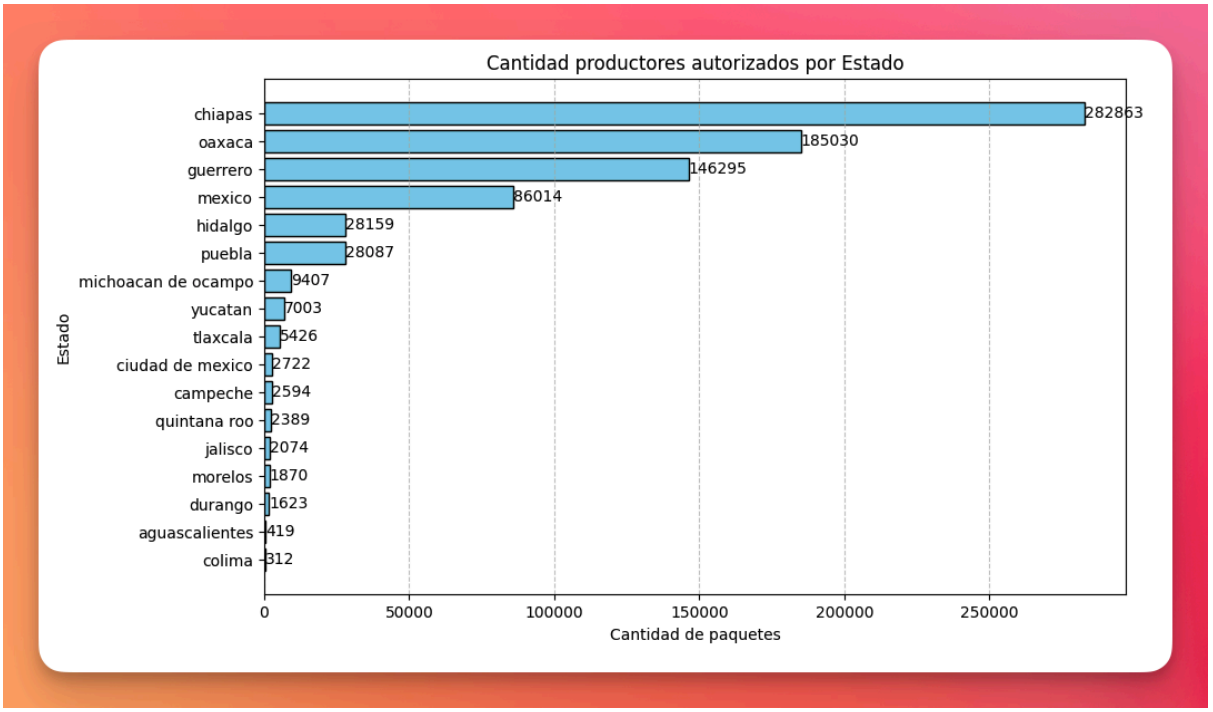
## 4.1. Variables numéricas.



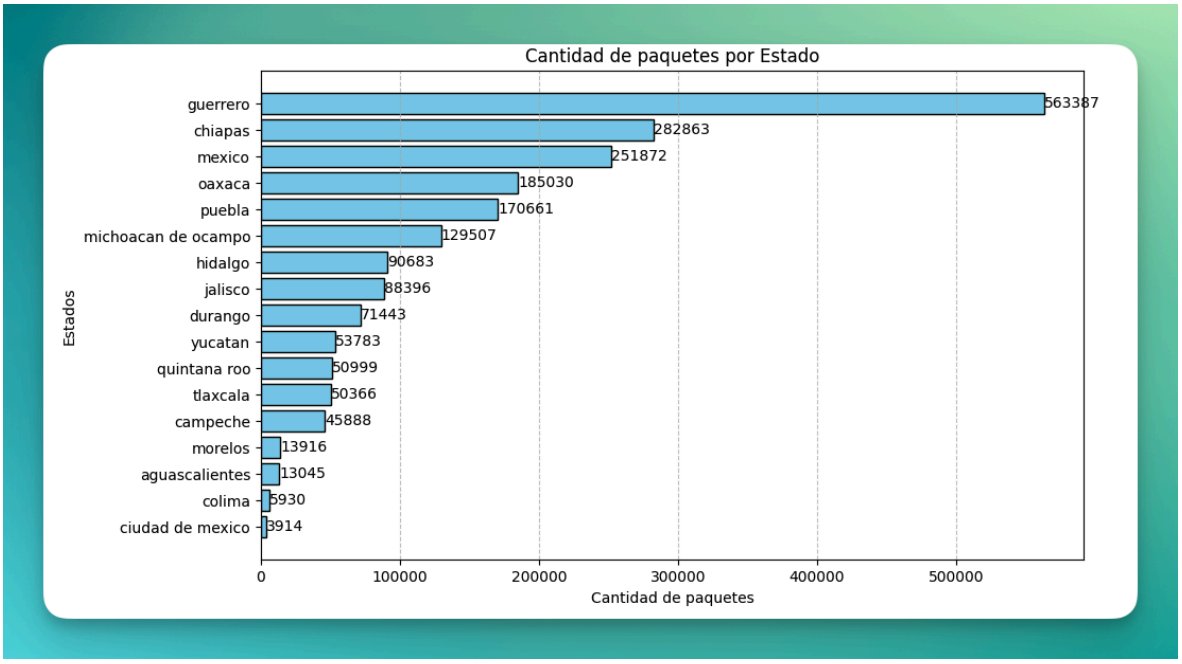

En este caso sólo disponemos de paquete como variable numérica, como ya comentamos antes la distribución es de uno y dos paquetes por productor autorizado.

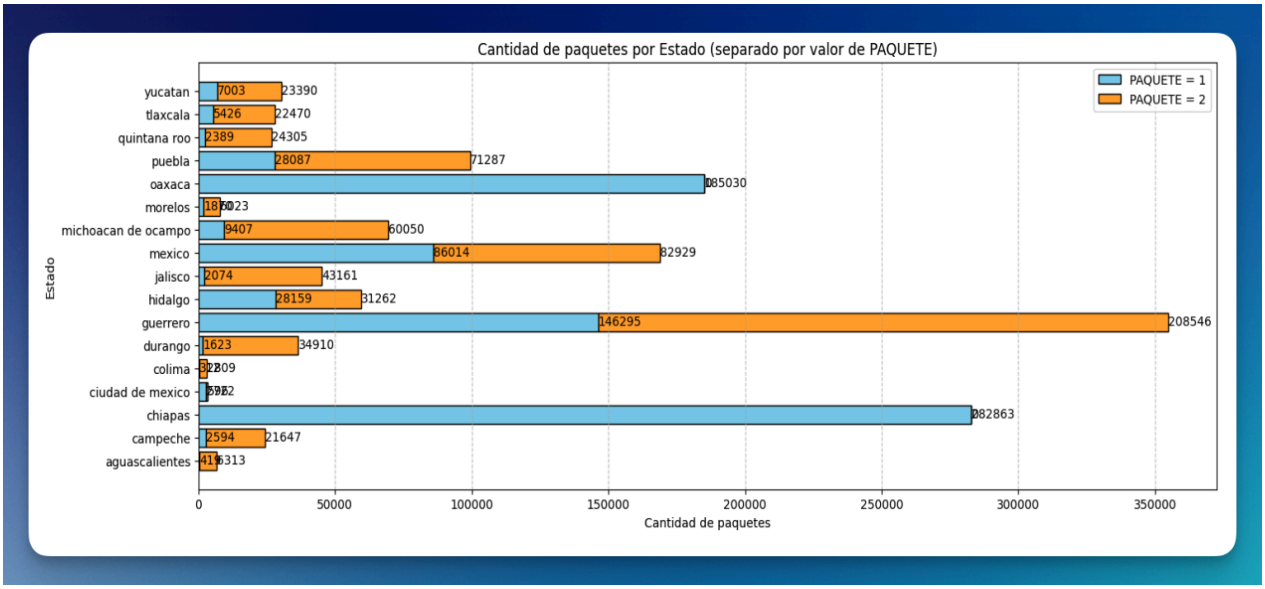
## 4.2. Variables categóricas.

## ESTADOS con más productores autorizados



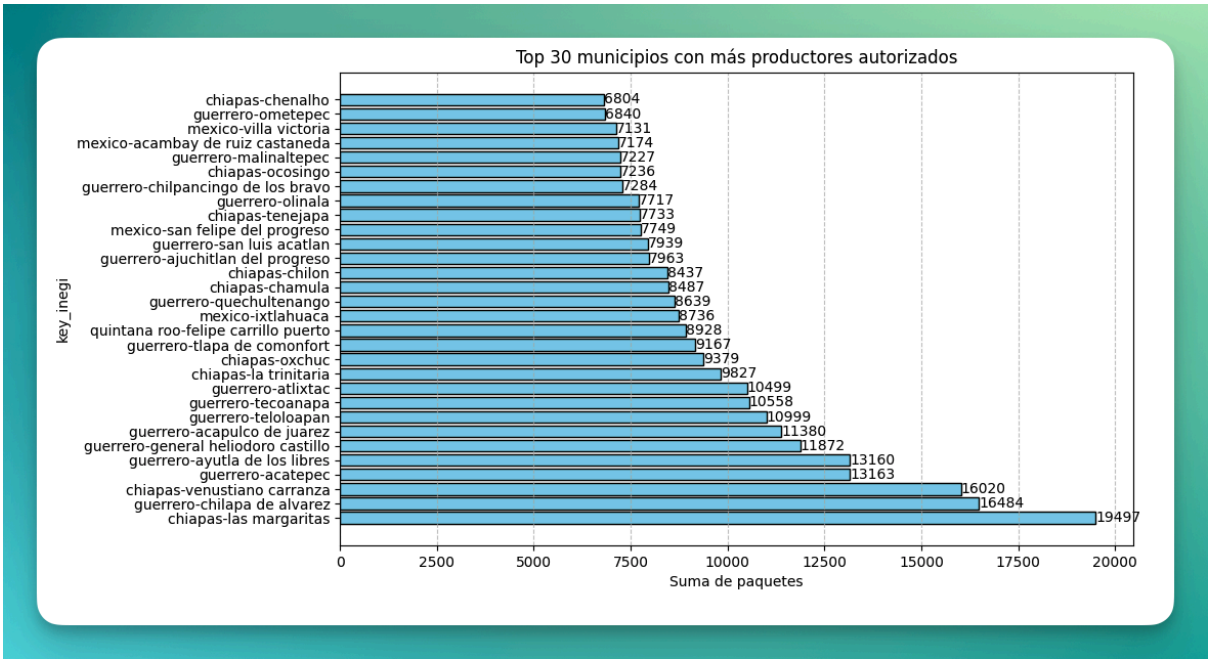

## ESTADOS con más paquetes productores autorizados






Como ya veníamos diciendo, el estado de Guerrero es el estado que más se repite y en el que se encuentran buena parte de los productores autorizados, cabe destacar que en CHIAPAS y OAXACA el número de productores autorizados también es elevado, sin embargo todos son autorizados para recibir un paquete. El siguiente estado con más productores autorizados es el de México.

Top 30 de Municipios con más productores autorizados




# 5. Extracción de conocimiento.

## 5.1. Principales hallazgos.

- Existen en México **1.525.720 productores autorizados**. (Teniendo en cuenta sólo 25 de las 44 bases de datos - ver link).
- Estos están distribuídos en todos los estados, sin embargo el 80% de ellos se concentran en los estados de:
    - Guerrero
    - Chiapas
    - México
    - Oaxaca
    - Puebla
    - Michoacán de Ocampo
- Los municipios con más productores autorizados son:
    - Las Margaritas (Chiapas)
    - Chilapa de Álvarez (Guerrero)
    - Venustiano Carranza (Chiapas)
- Los ACUSES son todos distintos, por lo que no se encuentran anomalías en esta variable.
- En el proceso recopilación de datos por parte de los funcionarios públicos se están cometiendo errores de escritura, lo que dificulta el proceso de auditoría posterior.
    - Se recomienda facilitar un aplicativo online con los campos de los municipios cerrados para evitar errores de captura de información.

## 5.2. Patrones, tendencias y relaciones significativas.

Teniendo en cuenta los datos a los que se han tenido acceso:

**Distribución Geográfica:**

- Estado de Guerrero: Es el estado con mayor número de productores autorizados, lo que indica una alta concentración de beneficiarios en esta región.
- Chiapas y Oaxaca: También muestran una alta concentración de productores autorizados, destacando como áreas prioritarias para el programa.

**Frecuencia de Municipios:**

- Municipio de Las Margaritas: Es el municipio más común tanto en las bases de datos de productores autorizados como en los datos de INEGI, lo que sugiere una significativa participación de esta localidad en el programa.

**Asignación de Paquetes:**

- Número de Paquetes: La mayoría de los productores reciben uno o dos paquetes, con una notable homogeneidad en la distribución de estos insumos.
- Chiapas: Todos los productores en este estado reciben consistentemente un solo paquete, lo que podría indicar una política específica de distribución en esta región.

## 5.3. Siguientes pasos.

Como siguientes pasos se seguirá trabajando en la reconciliación de beneficiarios por municipios reconciliando los nombre según la nomenclatura de INEGI.