TRABAJO PRÁCTICO
LABORATORIO DE DATOS

Integrantes:

- Tobias Llop (871/22)
- Delfina Stabile (819/22)
- Felipe Luc Pasquet (1084/22)

RESUMEN:

En este trabajo realizamos distintos procesos sobre una fuente de datos correspondiente al Padrón de Operadores Orgánicos Certificados de la República Argentina, para así alcanzar a desarrollar una conclusión sobre la existencia de una posible relación entre el desarrollo de la actividad orgánica y la proporción de mujeres en establecimientos productivos en cada departamento de las provincias argentinas.

INTRODUCCIÓN:

El objetivo general del presente trabajo práctico es ordenar y limpiar datos del Padrón de las siguientes fuentes de datos: Padrón de Operadores Orgánicos Certificados de la República Argentina, Distribución geográfica de establecimientos productivos, Localidades de la Base de Asentamientos Humanos de la República Argentina y Diccionario de CLAE.

Para poder analizarlos y ver si existe una relación entre el desarrollo de la actividad de producción orgánica y la proporción de mujeres empleadas en establecimientos productivos (no necesariamente orgánicos) en cada departamento de las provincias argentinas
Para eso seguimos los siguientes pasos:

1. Analizamos la forma normal en la que se encontraban nuestras fuentes de datos
2. A través del método GQM, analizamos la calidad de nuestras fuentes de datos y tomamos decisiones para mejorar esta última.
3. Diseñamos un DER en el que se encuentran cada uno de los data frames como entidades que utilizaremos para arribar a nuestro objetivo. Definimos para cada una de las entidades, su clave primaria, sus claves foráneas y sus dependencias funcionales.
4. Importamos los datos de nuestras fuentes a nuestros data frames
5. Generamos distintos reportes acerca de nuestros data frames utilizando consultas de SQL
6. Analizamos nuestros data frames a través de distintas herramientas de visualización

Por último, buscamos una relación entre la cantidad de operadores orgánicos y la proporción de mujeres empleadas en establecimientos productivos de una misma provincia-departamento, a través de ajustes lineales, y llegamos a la conclusión de que no existe tal relación, ya que contamos con mucha dispersión de nuestros datos.

DECISIONES TOMADAS:

Cuando los rubros se indefinen, ponemos como clae2 el numero 999, que refiere a “otros sectores”
Como en la base de datos de “localidades_bahra” faltaban datos de algunos departamentos de tierra del fuego, decidimos completarla con la fuente de datos “departamentos”: https://datos.gob.ar/dataset/jgm-servicio-normalizacion-datos-geograficos/archivo/jgm_8.10

PROCESAMIENTO DE DATOS:

Fuentes primarias:
La fuente principal “padrón-de-operadores-orgánicos-certificados” no está en 1ra formal, pues el atributo “productos” no es atómico, contiene más de un producto en algunos casos, y esto se puede separar en otra lista.
La clave candidata podría ser la combinación de los atributos “razón social” y “establecimiento”.

La otra fuente principal “distribucion_establecimientos_productivos_sexo” . Está en 2da forma normal, ya que hay algunas subdependencias funcionales, que se podrían separar en otra tabla (clae2, clae6 y letra). Pero todos los atributos dependen de la clave única ID, y todos los atributos son atómicos.

Fuentes secundarias:

claedict está en 2da forma normal, ya que todos los atributos son atómicos y tienen dependencia funcional total de la clave clae6. Pero las descripciones de clae 3, clae2 y letra, dependen de esos atributos y generan dependencias transitivas a clae6, y podrían ser separadas en otras tablas.

localidades_bahra está en 2da forma normal, ya que todos los atributos tienen dependencia funcional total de la clave gid. Pero hay muchas subdependencias funcionales que generan dependencias transitivas, y podrían ser separadas en otras tablas .

Calidad de datos.
Para mejorar la calidad de nuestras fuentes de datos se llevaron a cabo distintas decisiones detalladas a continuación.

En la fuente claedict tomamos como datos críticos a letra, clae2 y clae2_descr, estos se encontraban con muchas ocurrencias repetidas, lo que afectaba a la relevancia de nuestros datos. Este problema no estaba asociado a nuestro modelo, sino a la instancia, ya que en la base de datos se asociaba a clae2 con clae6 lo que provocaba que ocurran estas repeticiones.

Para dar una medida concreta acerca de la magnitud de este problema utilizamos la metodología GQM.
Definimos como GOAL que no haya tuplas de letra, clae2 y clae2_descr repetidas.
Como QUESTION planteamos la siguiente pregunta: ¿Cuántas tuplas se repiten?.
Como METRIC utilizamos la proporción de tuplas repetidas con respecto a la cantidad total de tuplas, que a través de consultas de SQL pudimos ver que era de alrededor del 90,1%.
Viendo esta métrica, tomamos la decisión de no agregar las repeticiones de tuplas a nuestro dataframe.

En la fuente localidades_bahra tomamos como datos críticos a las columnas nombre_departamento, codigo_indec_departamento, nombre_provincia y codigo_indec_provincia. Estos datos también se encontraban con muchas ocurrencias

repetidas que afectaban a la relevancia de nuestros datos. Este problema no estaba asociado a nuestro modelo, sino a la instancia ya que se asociaban nuestros datos críticos con otra base de datos de asentamientos, lo que provocaba que ocurran estas repeticiones.

Para dar una medida concreta acerca de la magnitud de este problema utilizamos la metodología GQM.
Definimos como GOAL que no haya tuplas de nombre_departamento, codigo_indec_departamento, nombre_provincia y codigo_indec_provincia repetidas
Como QUESTION planteamos la siguiente pregunta: ¿Cuántas tuplas se repiten?.
Como METRIC utilizamos la proporción de tuplas repetidas con respecto a la cantidad total de tuplas, que a través de consultas de SQL pudimos ver que era de alrededor del 85,3%.
Viendo esta métrica, tomamos la decisión de no agregar las repeticiones de tuplas a nuestro dataframe.
La fuente distribucion_establecimientos_productivos_sexo no presenta problemas de calidad, ya que no posee ocurrencias repetidas ni datos faltantes.

Para la fuente padrón-de-operadores-orgánicos-certificados tomamos como datos críticos a las columnas establecimiento y razón social. La columna establecimiento presentaba gran cantidad de valores indefinidos a los que llamaba “NC”, esto afectaba a la completitud de nuestros datos. Esta problemática está asociada a la instancia ya que todos los productores que tienen valores “NC” en la columna categoria_desc figuran como comercializadores o elaboradores, lo que nos hace sospechar que seguramente se hayan unido distintas fuentes de datos.

Para dar una medida concreta acerca de la magnitud de este problema utilizamos la metodología GQM.
Definimos como GOAL que el dato de establecimiento esté completo.
Como QUESTION planteamos la siguiente pregunta: ¿Cuál es la proporción de productores orgánicos que tienen el dato correspondiente a establecimiento con “NC”?.
Como METRIC utilizamos la proporción de razón social con el dato de establecimiento vacío, que a través de consultas de SQL pudimos ver que era de alrededor del: 30.3%

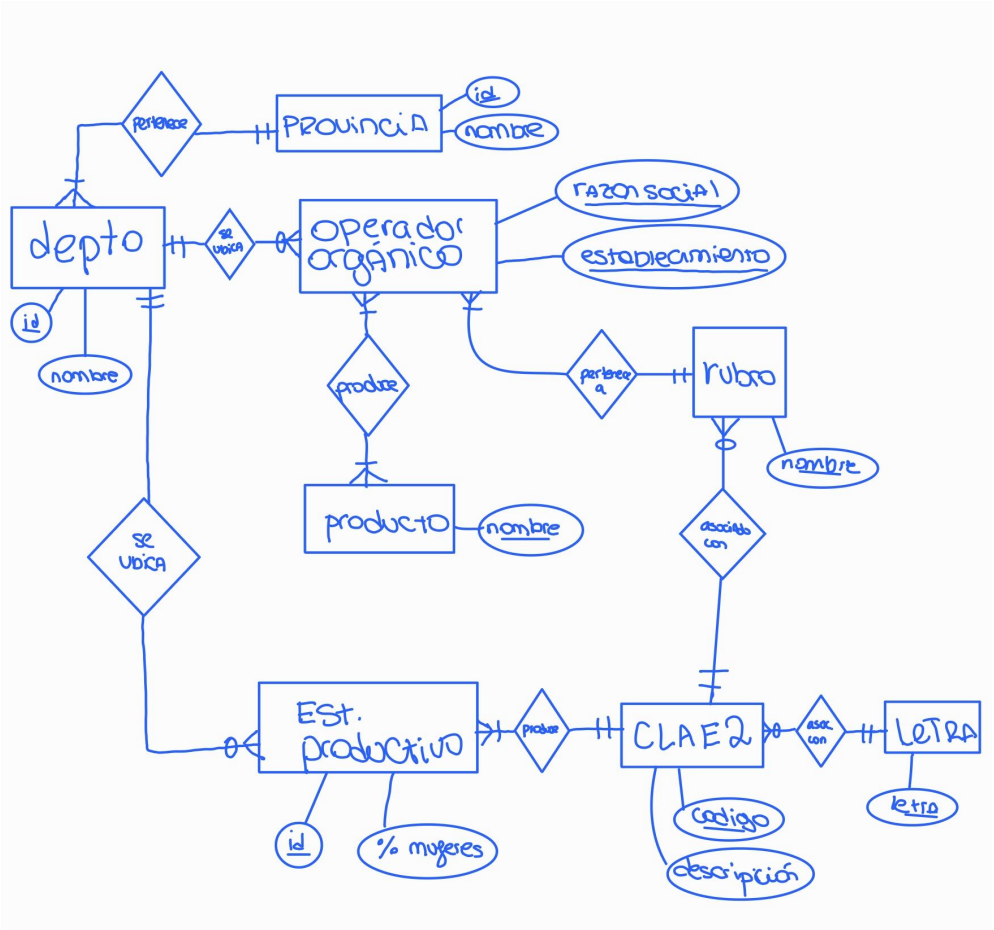
Viendo esta métrica, tomamos la decisión de asignarles a cada razón social con la columna NC vacía el mismo nombre de la razón social, ya que probablemente al ser elaboradores y comercializadores la mayoría de los productores orgánicos tengan una sola sucursal.

Otra problemática a la que nos enfrentamos fue a una gran cantidad de filas idénticas en nuestra tabla, que afectaban a la relevancia de nuestros datos. Este problema no estaba asociado a nuestro modelo, sino a la instancia.

Para dar una medida concreta acerca de la magnitud de este problema utilizamos la metodología GQM.
Definimos como GOAL que todas las filas de la tabla sean distintas,
Como QUESTION planteamos la siguiente pregunta: ¿Cuál es la proporción de filas idénticas en nuestra tabla?
Como METRIC utilizamos la proporción de filas repetidas, que a través de consultas de SQL pudimos ver que era de alrededor del: 4,3%

Como la proporción de filas idénticas era menor al 10% decidimos eliminar estos datos.

Luego de la limpieza de nuestros datos, creamos data frames vacíos para cada una de las entidades de nuestro DER (teniendo en cuenta si las relaciones eran uno a uno, uno a muchos o muchos a muchos), para así agregarle nuestros datos de interés, a continuación detallamos a cada data frame con su primary key, sus foreign key y sus dependencias funcionales:



CLAE2(Código, descripción, letra) letra depende de Letra(letra)
Op_Organico(razón social, establecimiento, id_depto, nombre_rubro) id_depto depende de Depto(id) y nombre_rubro depende de Rubro(nombre)
Depto(id, nombre, id_prov) id_prov depende de provincia(id)
provincia(id, nombre)
Producto(nombre)
Rubro(nombre, cod_clae2) cod_clae2 depende de Clae2(código)
Est_Productivo(id, porc_mujeres, id_depto, cod_clae) id_depto depende de Depto(id) y cod_clae depende de Clae2(codigo)
Letra(letra)
Oorg_produce(razon_social, establecimiento, nombre_prod) razón_social depende de op_organico(razón social), establecimiento depende de Op_organico(establecimiento) y nombre_prod depende de producto(nombre)

Importación de datos

A nuestro dataframe clae2 le importamos los datos de las columnas clae2, clae2_desc y letra de nuestra fuente Diccionario CLAE.

A nuestro dataframe letra le importamos los datos de la columna letra de nuestra fuente Diccionario CLAE usando un SELECT DISTINCT en sql.

A nuestro dataframe est_productivos le importamos los datos de las columnas ID, proporcion_mujeres, in_departamentos y clae2 de nuestra fuente Distribución geográfica de los establecimientos productivos,

A nuestro dataframe depto le importamos los datos de las columnas codigo_indec_departamento, codigo_indec_provincia y nombre_departamento de nuestra fuente Localidades de la Base de Asentamientos Humanos de la República Argentina.

A nuestro dataframe op_organicos le importamos los datos de las columnas razón social, establecimiento, rubro e id de nuestras fuentes Localidades de la Base de Asentamientos Humanos de la República Argentina y Padrón de Operadores Orgánicos Certificados.

A nuestro dataframe provincia le importamos los datos de las columnas codigo_indec_provincia y nombre_provincia de nuestra fuente Localidades de la Base de Asentamientos Humanos de la República Argentina.

A nuestro dataframe op_organico_produce le importamos los datos de las columnas razón social, establecimiento y productos de nuestra fuente Padrón de Operadores Orgánicos Certificados.

A nuestro dataframe producto le importamos los datos de la columna productos de nuestra fuente Padrón de Operadores Orgánicos Certificados..

A nuestro dataframe rubro_clae2 le asignamos a mano una clae2 a cada rubro de la columna rubro de nuestra fuente Padrón de Operadores Orgánicos Certificados.

ANÁLISIS DE DATOS:

A través de consultas SQL pudimos llegar a las siguientes conclusiones:

- Observamos en qué provincia se produce cada producto producido por operadores orgánicos. Ver anexo 1
- Pudimos ver que el CLAE más común es el que tiene como código al 47 y como descripción: Comercio al por menor excepto autos y motos. Este CLAE se encontraba con 130259 apariciones en nuestro data frame Est_productivo.
- Observamos que el producto más producido es la caña de azúcar y se produce en los departamentos-provincia listados en el anexo 2.
- Vimos que existen 378 departamentos que no presentan operadores orgánicos y se encuentran listados en el anexo 3.
- Observamos la cantidad de establecimientos productivos y emprendimientos orgánicos que posee cada provincia-departamento. ver anexo 4

- Analizamos la tasa promedio de participación de mujeres a nivel nacional: `0.3346911181;` y la tasa promedio de participación de mujeres en cada provincia. Luego analizamos qué provincias están por encima del promedio (y cuáles por debajo) en una tabla. (ver anexo 5)
- Calculamos el desvío del promedio de participación de mujeres en las provincias y fue del: 0.04076714668

Mediante herramientas de visualización realizamos los siguientes gráficos:

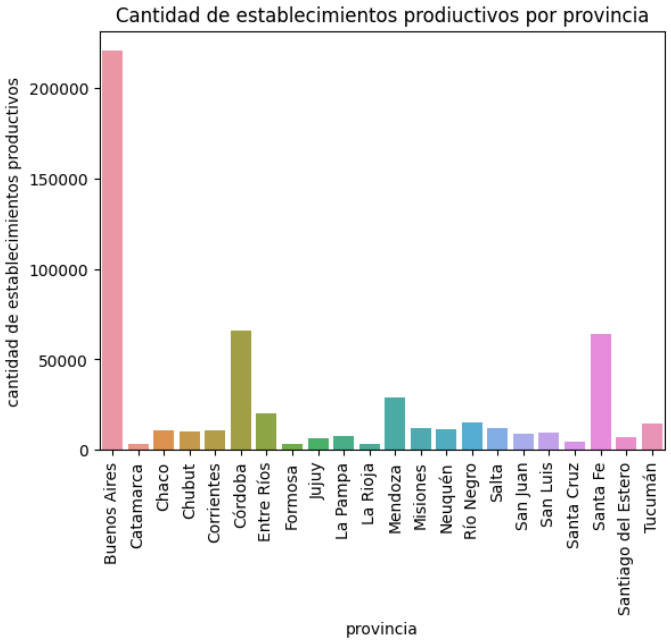


Gráfico 1: Histograma de la cantidad de establecimientos productivos por provincia

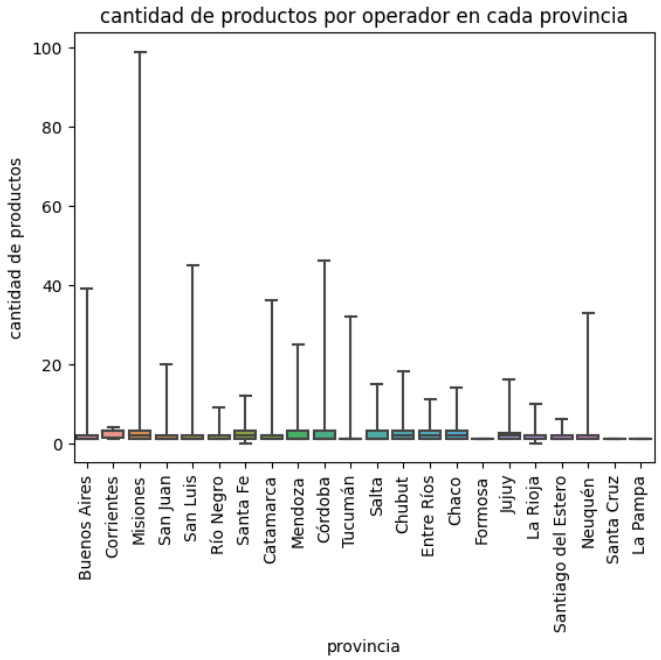


Grafico 2: boxplot que demuestra la cantidad de productos por cada operador organico en cada una de las provincias

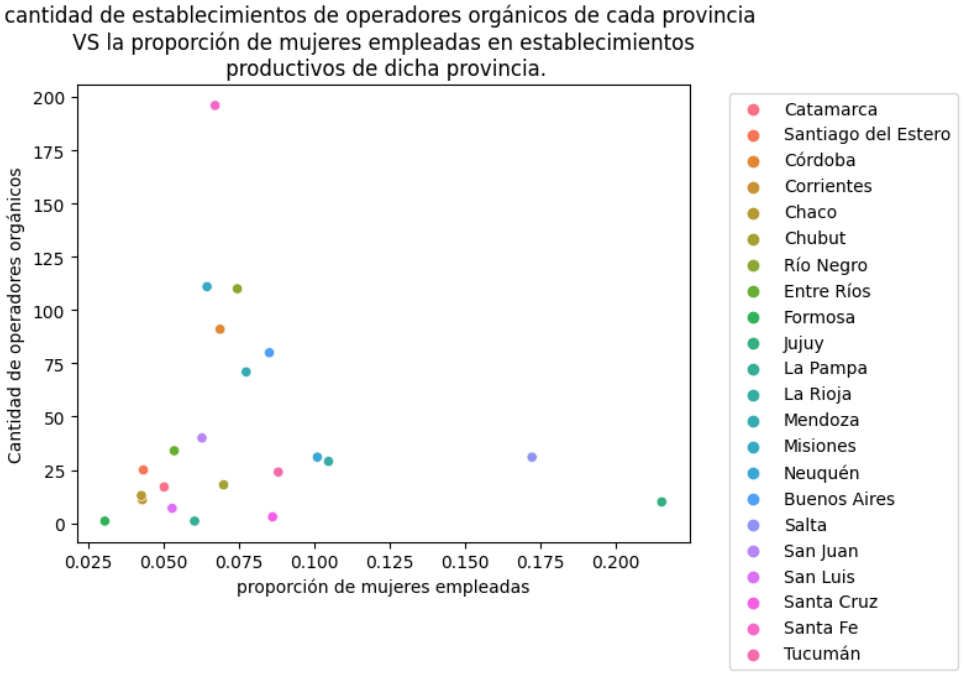


Gráfico 3: Relación entre cantidad de establecimientos de operadores orgánicos certificados de cada provincia y la proporción de mujeres empleadas en establecimientos productivos de dicha provincia para una misma letra de CLAE2.

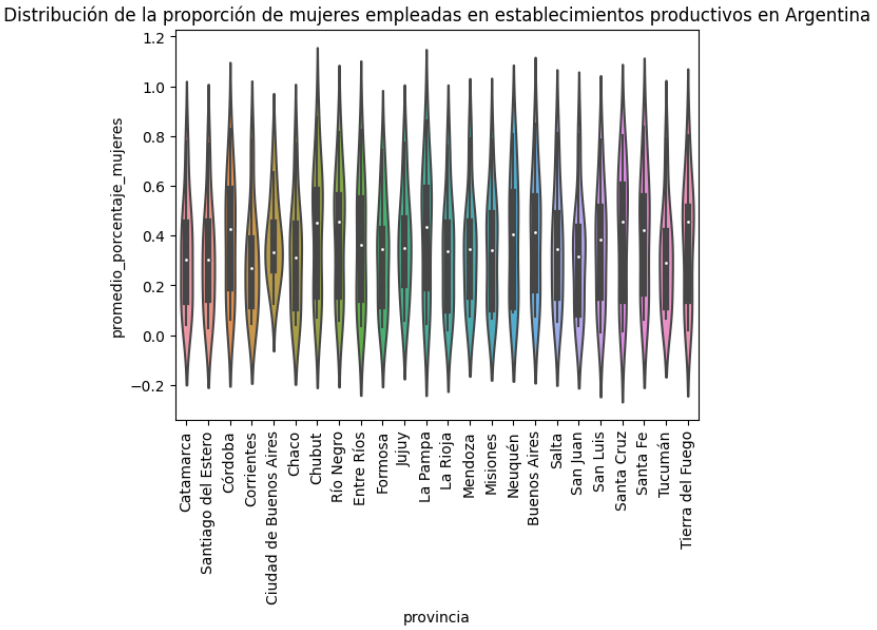
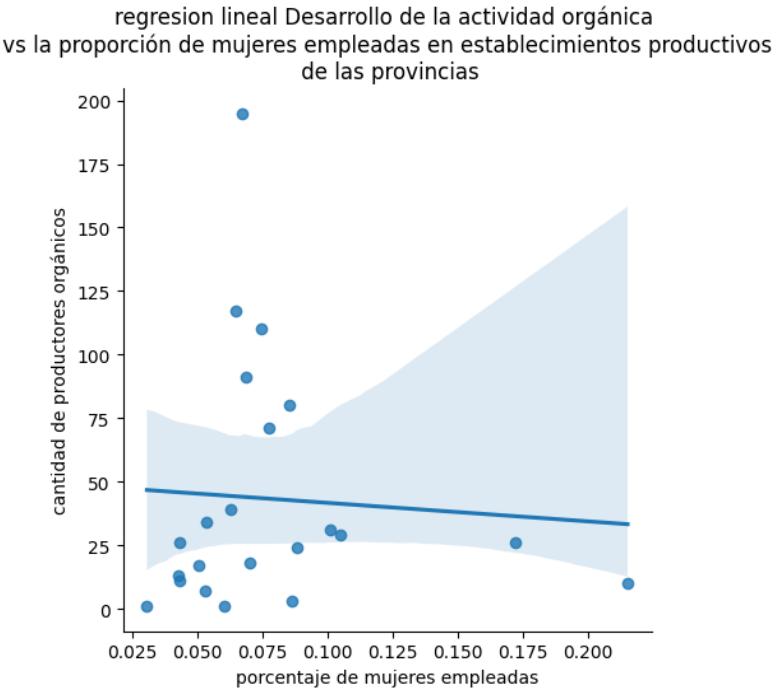
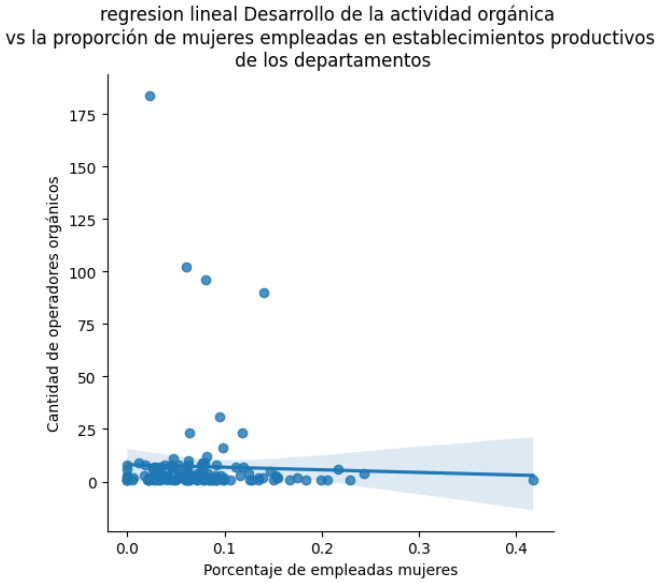


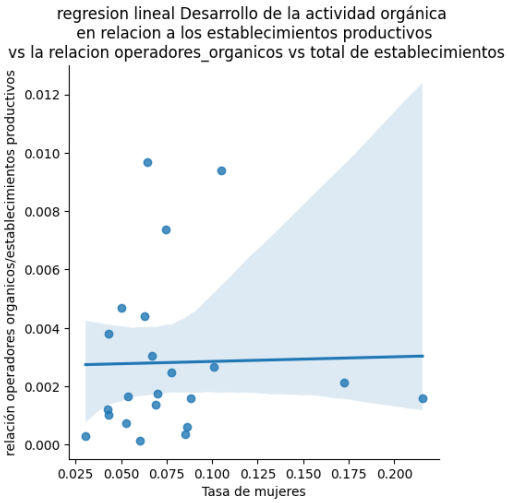
Gráfico 4: violin plot de la distribución de la proporción de mujeres trabajando en establecimientos productivos, en cada una de las provincias.

CONCLUSIONES:

Tras la limpieza de nuestros datos y su posterior análisis y visualización a través de consultas de SQL y la utilización de las librerias seaborn y matplotlib para realizar distintos gráficos. Buscamos una relación entre la cantidad de productores orgánicos y la proporción de mujeres en establecimientos productivos con una misma letra de CLAE para cada provincia y para cada departamento. Diseñamos estas tablas a través de consultas SQL y las ajustamos linealmente en los siguientes gráficos:







En los gráficos no se alcanza a ver una relación entre ambas variables ya que el ajuste deja muchos datos afuera del ajuste, ya que los datos están muy dispersos.
Debido a esto, finalmente concluímos que no existe una relación entre el desarrollo de la actividad orgánica y la proporción de mujeres en establecimientos productivos.