DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 1

Análisis de la relación entre el desarrollo de la actividad de producción orgánica y la proporción de mujeres empleadas en establecimientos productivos en cada departamento de las provincias argentinas.

---

21 de octubre de 2023 — Laboratorio de Datos

GRUPO LMJ

Integrante	LU	Correo electrónico
Alamo, Malena Sol	1620/21	malusalamo@gmail.com
Laria Guaza, Jeremias	1329/21	jeremiaslaria7@gmail.com
Tag, Lucio	876/22	luciotag2011@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

RESUMEN

En el trabajo se desea analizar la 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 el análisis se utilizaron las siguientes fuentes:

- Operadores orgánicos. Padrón de Operadores Orgánicos Certificados, cuyo responsable es la Dirección de Agroalimentos - Producción Orgánica
- Establecimientos productivos. Distribución geográfica de los establecimientos productivos. Esta fuente de datos contiene coordenadas de los establecimientos productivos (no necesariamente orgánicos) con su respectiva actividad económica, su nivel de empleo, jurisdicción y proporción de mujeres.
- Localidades. Localidades de la Base de Asentamientos Humanos de la República Argentina
- CLAE. Diccionario de CLAE, cuyo responsable es el Ministerio de Desarrollo Productivo. Unidad Gabinete de Asesores. Dirección Nacional de Estudios para la Producción (CEP XXI).

Para el desarrollo del análisis de datos se optó por el diseño de una base de datos relacional, la importación de sus correspondientes datos y luego las consultas y el análisis para responder determinadas preguntas utilizando SQL y visualizando con diferentes librerias de python.

Se detalla en los siguientes apartados la exposición del trabajo.

INTRODUCCIÓN

Se toma como objetivo general determinar distintas relaciones entre el desarrollo de la producción orgánica y los distintos establecimientos productivos de la República Argentina. Deseamos analizar los detalles de la producción orgánica. Son de nuestro interés cuestiones como los productos producidos por operadores orgánicos, la distribución geográfica, la proporción de mujeres, la cantidad de establecimientos, etc.

Además, se pretende encontrar las relaciones entre los mismos atributos. Como objetivo principal debemos encontrar la relación entre la producción orgánica y la proporción de mujeres dentro de los establecimientos productivos.

Para desarrollar el objetivo se siguieron los siguientes pasos:

- Diseño de la base de datos a través de un Modelo Entidad Relación. A partir de las preguntas que queremos contestar, diseñamos las tablas con la información que necesitabamos. Las mismas las diseñamos en 3FN.
- Importación y limpieza de datos: a partir de las fuentes primarias y secundarias importamos y realizamos la limpieza de toda la información.

- Respuesta a las preguntas con SQL: a través de la unión y combinación de las tablas ya limpias, encontramos las respuestas que eran de nuestra necesidad.
- Visualización: visualizamos la información utilizando Pandas, Matplotlib y Seaborn para seguir encontrando relaciones y entendiendo la información.

DECISIONES TOMADAS

Durante el avance del diseño de las tablas que utilizamos como información nos encontramos con diferentes situaciones que requerian la toma de decisiones. Dejamos detalladas las mismas:

Departamentos

Para solucionar el problema de que hay varios departamentos con el mismo nombre por provincia, le cambiamos el nombre a los departamentos a ‘departamento’ + ‘provincia’

Productos

- Se consideran los siguientes productos como el único producto PASTURA: PASTURAS, PASTIZAL NATURAL, PASTO LIMON, PASTOS NATURALES E IMPLANTADOS
- Se considera el producto EXTRACCIÓN DE MIEL como MIEL
- Se consulderan los siguientes productos como el único producto FRUTOS: FRUTALES, FRUTAS TROPICALES, FRUTICULTURA

Operadores organicos certificados:

Para responder si existen departamentos que no presentan Operadores Organicos Certificados, tomamos los Operadores Organicos Productores que son los que poseen establecimiento.

Desvio Estandar:

Para calcular el desvio estandar de una provincia, calculamos el promedio de los desvio estandar de cada uno de los departamentos de dicha provincia.

Cantidad productos por operador:

Para la visualizar la cantidad de productos por operador se tomaron los operadores productivos, sin tener en cuenta elaboradores y comercializadores.

Emprendimientos organicos:

Asumimos emprendimientos organicos como operadores organicos productivos.

PROCESAMIENTO DE DATOS

Veamos en que forma normal se encontraban las fuentes de datos originales:

- Operadores Organicos..

Vemos que no se encuentra en 1FN ya que productos no es un atributo atomico.

pais_id	pais	provinci...	provincia	departa...	localidad	rubro	productos
32	ARGEN...	6	BUENO...	BARAD...	INDEFIN...	AGRICU...	SOJA
32	ARGEN...	6	BUENO...	BARAD...	INDEFIN...	AGRICU...	GIRASOL, MIJO, RESIDUO VEGETAL

En la segunda tupla podemos ver que puede ser descompuesta en 3 tuplas en las que cada una tenga uno de los 3 productos.
Como no se encuentra en 1FN entonces tampoco se encuentra en ninguna de las formas normales vistas en la materia.

- Establecimientos Productivos.

Esta tabla se encuentra en 1FN ya que todos sus atributos son atomicos y no posee relaciones dentro de relaciones ni relaciones como valores de atributos.
Tambien se encuentra en 2FN ya que la PK esta compuesta de un solo atributo que es ID, por lo que todos los atributos no primos dependen enteramente de la PK.
Podriamos decir que no esta en 3FN ya que pareceria haber una DF transitiva de ID -> Departamento, Departamento -> Provincia, ya que es logico pensar que el departamento nos implica la provincia. Pero esto no es asi ya que por ejemplo tenemos el departamento Capital que se repite en varias provincias, durante el desarrollo del TP cambiamos esto y a los departamentos que se llamaban Capital les agregamos el nombre de la provincia ya que nos generaba problemas, esto igualmente no soluciona el problema porque siguen habiendo departamentos que se llaman igual en distintas provincias.
No se encuentra en 3FN ya que ID->Provincia y Provincia->provincia_id.

- Localidades.

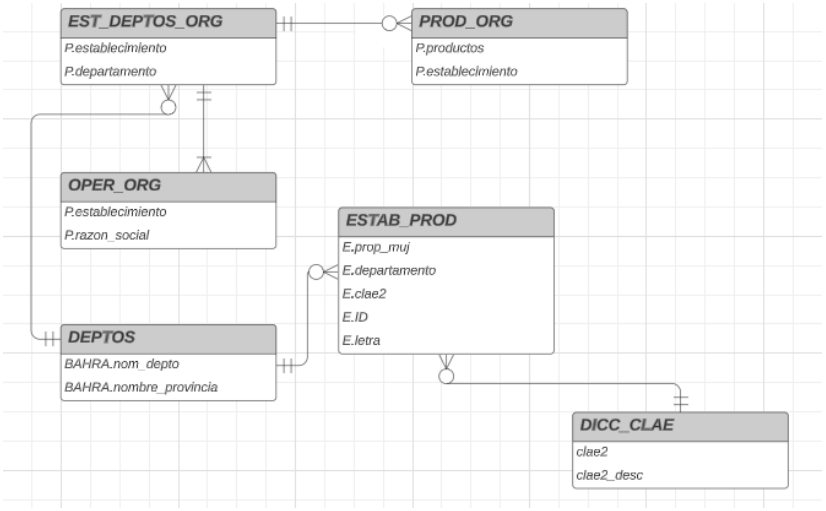
Esta en 1FN ya que todos sus atributos son atomicos.
La PK esta compuesta de un solo atributo que es gid, por lo que esta en 2FN.
Como gid->nombre_provincia y nombre_provincia->codigo_indec_provincia, la tabla no se encuentra en 3FN.

- CLAE.

Todos los atributos son atomicos y no posee relaciones dentro de relaciones ni relaciones como valores de atributo, por ende esta en 1FN.
Nuestra PK esta compuesto por un solo atributo que es clae6, por lo tanto esta en 2FN.
Como clae6 -> clae3, y clae3 -> clae2, tenemos una DF transitiva, por ende no esta en 3FN.

Para aumentar la calidad de los datos el problema principalmente tuvimos que solucionar los problemas que habia con los nombres de los establecimientos y los nombres de los productos, donde en establecimientos por ejemplo tuvimos que arreglar los nombres que contenian caracteres como °, ü ó ñ. Y para productos tambien tuvimos que corregir nombres con estos caracteres y aquellos nombres que contenian espacios al final o al principio del string.
Para las demas tablas no hubo mayores problemas, tuvimos que corregir las razones sociales que terminaban con S.A, S.R.L, etc. ya que algunas estaban como S.A y otras como S.A. por ejemplo.

DIAGRAMA ENTIDAD RELACION:



Pasamos a explicar brevemente las relaciones entre las tablas:

- EST_DEPTOS_ORG, PROD_ORG

Tenemos una relacion uno a muchos ya que a cada establecimiento de la tabla de establecimientos y departamentos organicos le corresponde muchos productores o ninguno, y a cada productor le corresponde un y solo un establecimiento.

- EST_DEPTOS_ORG, OPER_ORG.

Por cada establecimiento organico tenemos uno o mas operadores organicos, ya que podria ser un establecimiento compartido (suceso que ocurre en la tabla), y cada operador organico tiene un y solo un establecimiento.

- EST_DEPTOS_ORG, DEPTOS.

Esta relacion surge de la necesidad de poder saber a que provincia pertenece cada departamento, para cada departamento tenemos una y solo una provincia a la que pertenece y cada departamento puede tener muchos establecimientos organicos o no tener ninguno.

- DEPTOS, ESTAB_PROD.

Cada departamento tiene muchos establecimientos productivos o ninguno, y cada establecimiento productivo pertenece a un solo departamento.

- ESTAB_PROD, DICC_CLAE

Cada establecimiento productivo tiene un clae2 asociado que nos describe la actividad de ese establecimiento. Cada establecimiento le corresponde un unico clae2 necesariamente, y pueden haber clae’s2 que tengan muchos establecimientos productivos o que no tenga ninguno.

ANÁLISIS DE DATOS

i) Para cada producto (producido por un productor orgánico) detallar en qué provincias se produce. El orden del reporte debe respetar la cantidad de provincias en las cuales se produce dicho producto (de mayor a menor). En caso de empate, ordenar alfabéticamente por nombre de producto.

Lo primero que hicimos fue unir nuestras tablas df_prod_org y df_est_deptos_org para en una misma tabla tener PRODUCTO-ESTABLECIMIENTO-DEPARTAMENTO, luego la unimos con la tabla df_deptos para tener la tabla PRODUCTO-PROVINCIA. Después las agrupamos por producto y provincia para hacer un conteo. Finalmente unimos esas dos tablas para ordenar prod_prov según cant_prov_prod.

	productos	nombre_provincia
0	VID	Río Negro
1	VID	Santa Fe
2	VID	San Juan
3	VID	Mendoza
4	VID	La Rioja
..	...	...
568	VACAS DE TAMBO	Buenos Aires
569	VD	La Rioja
570	YERBA MATE	Misiones
571	ZAPALLITO	Mendoza
572	ZUCCINI	Mendoza

[573 rows x 2 columns]

ii) ¿Cuál es el CLAE2 más frecuente en establecimientos productivos? Mencionar el Código y la Descripción de dicho CLAE2.

Primero agrupamos df_estab_prod según que clae tenia y contamos la cantidad de establecimientos. Despues la unimos con df_dicc_clae para que aparezca la descripción y la ordenamos según cantidad. Finalmente concluimos en que el código de clae mas frecuente en establecimientos productivos es **47** y su descripción es "**comercio al por menor excepto auto".**

	codigo_clae	descripcion_clae
0	47	Comercio al por menor excepto autos y motos

iii) ¿Cuál es el producto más producido (que lo producen más establecimientos de operadores orgánicos)?¿Qué Provincia-Departamento los producen?

Aca empezamos agrupando df_prod_org por producto y contando la cantidad de establecimientos, para luego unirla con prod_est_dep(tabla de la primer consulta) con departamentos para agregar la provincia y filtramos por manzana que ya desde un primer momento averiguamos que era el producto mas producido.

```
Producto mas producido
  productos  cantidad_establecimientos
0  MANZANAS                       153

Departamentos donde se produce manzana
   productos     departamento nombre_provincia
0   MANZANAS        SAN PEDRO     Buenos Aires
1   MANZANAS       LAS FLORES     Buenos Aires
2   MANZANAS  CORONEL DORREGO     Buenos Aires
3   MANZANAS     TRES ARROYOS     Buenos Aires
4   MANZANAS        ESCALANTE           Chubut
5   MANZANAS     GENERAL ROCA          Córdoba
6   MANZANAS        SAN PEDRO            Jujuy
7   MANZANAS       SAN CARLOS          Mendoza
8   MANZANAS       SAN RAFAEL          Mendoza
9   MANZANAS        SAN PEDRO         Misiones
10  MANZANAS      CONFLUENCIA          Neuquén
11  MANZANAS     GENERAL ROCA        Río Negro
12  MANZANAS           EL CUY        Río Negro
13  MANZANAS       SAN CARLOS            Salta
```

iv)¿Existen departamentos que no presentan Operadores Orgánicos Certificados? ¿En caso de que sí, cuántos y cuáles son?

Si, aca lo primero que hicimos fue agarrar a los departamentos con operadores organicos certificados y luego hicimos un codigo para que los departamentos que esten en la tabla de departamentos y que no este en la query anterior, no vayan a tener OOC´s.

```
              nombre_departamento
0                       Arrecifes
1                 Coronel Suárez
2              General San Martín
3                         Quilmes
4                          Ambato
..                            ...
405                      Iglesia
406                  Corpen Aike
407  Belgrano Santiago del Estero
408                         Choya
409       Islas del Atlántico Sur

[410 rows x 1 columns]
```

v) ¿Cuál es la tasa promedio de participación de mujeres en cada provincia?¿Cuál es su desvío? En cada caso, mencionar si es mayor o menor al promedio de todo el país

Primero agrupamos los Est_prod por departamento y calculamos la media de la proporción de mujeres y el desvio, luego la unimos con df_deptos para poder agregar cada provincia con su media y su desvio, luego calculamos la media total del pais para ver si estaba por encima o por debajo de dichas provincias.

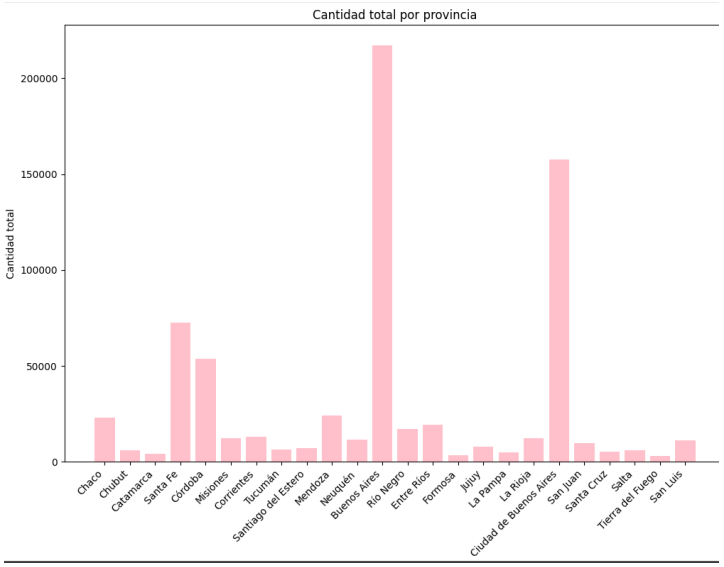
	nombre_provincia	media_mujeres	desvio	desvio_mayor_a_media
0	Buenos Aires	0.319534	0.400008	True
1	Catamarca	0.213999	0.330590	True
2	Chaco	0.208017	0.332590	True
3	Chubut	0.231796	0.348631	True
4	Ciudad de Buenos Aires	0.358488	0.403246	True
5	Corrientes	0.234346	0.352993	True
6	Córdoba	0.303628	0.386657	True
7	Entre Ríos	0.269771	0.386213	True
8	Formosa	0.213995	0.367264	True
9	Jujuy	0.281006	0.338167	True
10	La Pampa	0.264174	0.395328	True
11	La Rioja	0.233162	0.333664	True
12	Mendoza	0.260702	0.363363	True
13	Misiones	0.256350	0.369010	True
14	Neuquén	0.388901	0.400314	True
15	Río Negro	0.293935	0.383079	True
16	Salta	0.246211	0.337622	True
17	San Juan	0.204834	0.339861	True
18	San Luis	0.270673	0.376157	True
19	Santa Cruz	0.388455	0.409878	True
20	Santa Fe	0.308529	0.404215	True
21	Santiago del Estero	0.224748	0.325585	True
22	Tierra del Fuego	0.382618	0.406430	True
23	Tucumán	0.236733	0.342981	True

vi) Mostrar por cada provincia-departamento cuántos establecimientos productivos y cuántos emprendimientos orgánicos posee

Para este primero unimos entre df_estab_prod y df_deptos para agregar su dicha provincia y luego las agrupamos por departamento y provincia. Luego hicimos una union entre df_estab_prod y df_deptos para agregar la provincia y lo agrupamos por departamento y provincia. Finalmente unimos ambas tablas.

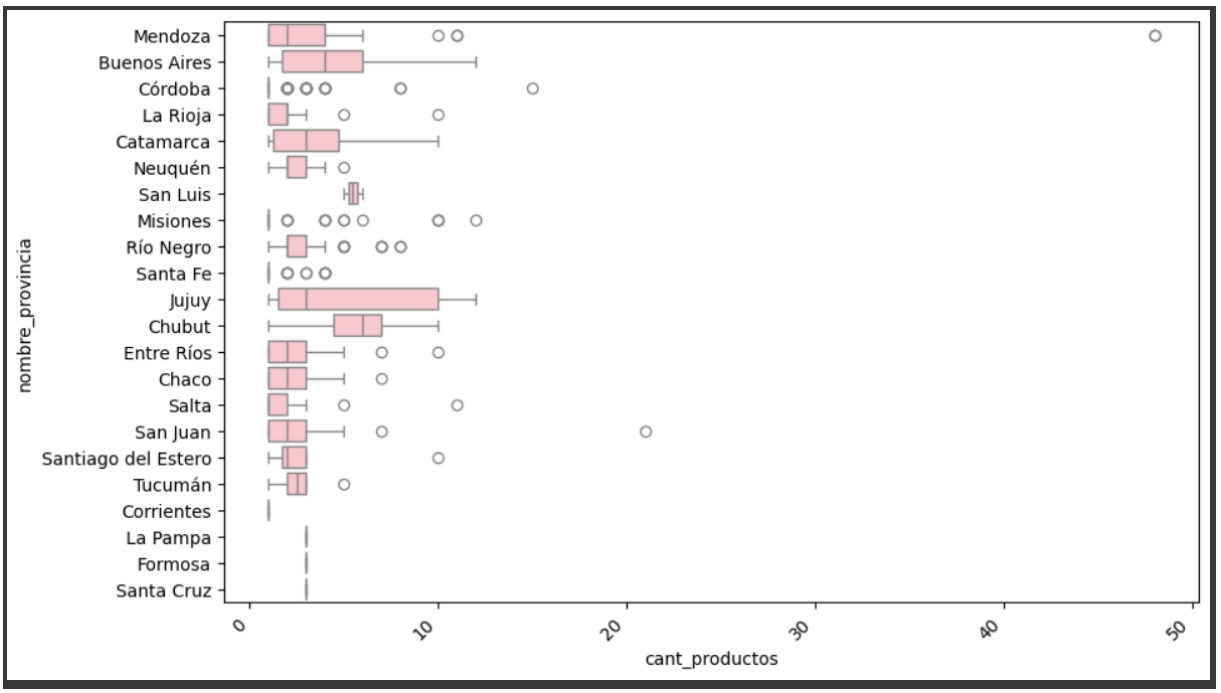
Visualización

i)Cantidad de establecimientos productivos por provincia
Realizamos un gráfico de barras para poder observar la cantidad de establecimientos productivos por provincia. Observamos que las provincias que más tienen establecimientos son Buenos Aires (incluyendo la Ciudad de Buenos Aires), Santa Fe y Córdoba



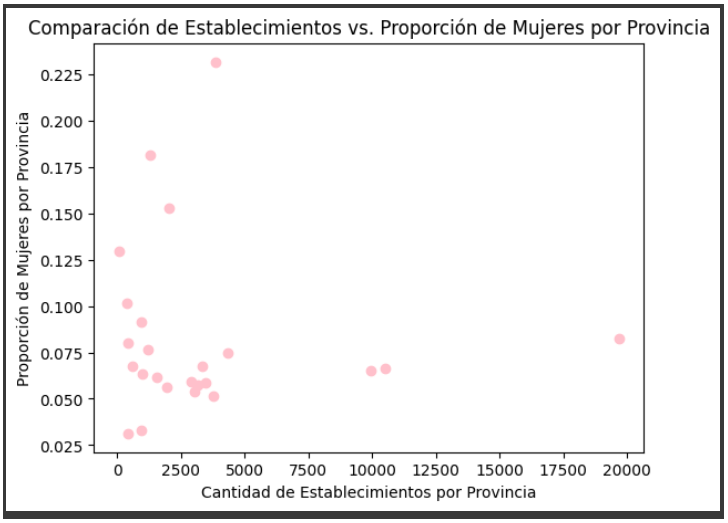
ii) Boxplot, por cada provincia, donde se pueda observar la cantidad de productos por operador

A partir del boxplot podemos analizar que la provincia Jujuy es la que tiene más variedad de cantidad de productos por operador. Observamos el valor más atipico en Mendoza, con un operador que tiene aproximadamente 50 productos.

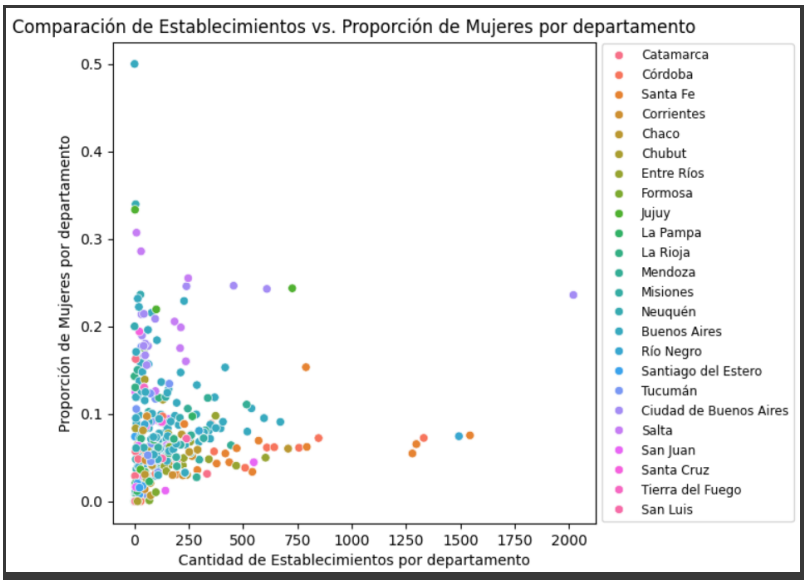


iii) Relación entre cantidad de establecimientos de operadores orgánicos certificados de cada provincia y la proporción de mujeres empleadas en establecimientos productivos de dicha provincia. Para este punto deberán generar una tabla de equivalencia, de manera manual, entre la letra de CLAE y el rubro de del operador orgánico.

Primero analizamos la comparación de cantidad de establecimientos vs proporción de mujeres por provincia con un scatterplot. Vemos que no presentan relación

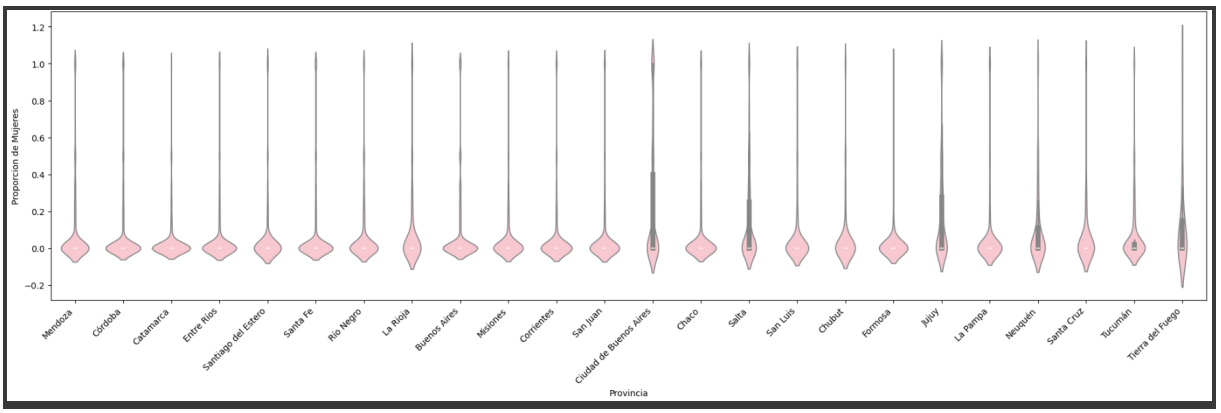


Para hacer un análisis un poco más grande e intentar encontrar alguna relación, realizamos la comparación entre la cantidad de establecimientos y la proporcion de mujeres por departamento, segmentando según la provincia. Este gráfico tampoco nos muestra relación.



iv) ¿Cuál es la distribución de los datos correspondientes a la proporción de mujeres empleadas en establecimientos productivos en Argentina? Realicen un violinplot por cada provincia. Mostrarlo en un solo gráfico.

Realizamos a través de un violinplot la distribución de la proporción de mujeres empleadas según provincia. Vemos que la distribución se acumula en valores bajos.



Conclusiones

Durante el análisis no se encontró relación entre el desarrollo de la actividad orgánica y la proporción de mujeres empleadas en establecimientos productivos de las provincias.

Se puede observar que la proporción de mujeres en establecimientos productivos es baja, independientemente de la provincia y departamento (aunque hay algunas excepciones).

Se puede hacer también la observación de la calidad de datos de las fuentes utilizadas: para la columna establecimiento de la fuente establecimientos.csv y para la columna productos de la fuente padron.csv se debio realizar una limpieza exhaustiva. Esto puede generar diferentes errores a la hora de realizar el análisis de la información.

Otro problema que podemos encontrar es la falta de relación entre la letra correspondiente al rubro de los Operadores Orgánicos y dicha letra en Establecimientos Productivos. Al filtrar establecimientos productivos por la letra A, obtenemos una cantidad extremadamente más grande que productores orgánicos (tenemos aproximadamente 1300 Operadores Orgánicos y el filtro nos devuelve 69000 establecimientos productivos aparentemente orgánicos). Debido a esto consideramos que tal vez no es apropiado analizar la relación utilizando las 69000 filas asumiendo que son establecimientos orgánicos. Puede surgir de esta situación el resultado de la falta de relación.

Además, consideramos que juega un papel importante en el análisis el conocimiento que se tiene sobre el tópico a desarrollar. En este caso, no poseíamos ningún manejo sobre el tema del trabajo (Establecimientos productivos y Operadores Orgánicos en la Argentina). Esto puede ocasionar que pasen desapercibidos ciertos detalles que tal vez una persona que tiene profundos conocimientos en el tema sepa entender más rápido.